Erasmus School of Economics

Thesis

To obtain the academic degree of

Master of Science in Economics & Business

(Major in Marketing)

Sub topic identification within texts

Author: Vincent van Wingerden

Email address: vincentvanwingerden@outlook.com

Supervisor: dr. F. Florian Deutzmann

Study Program: Business Economics

Specialization: Marketing

Date: 04-01-16

Abstract

This Thesis explores the possibility to use machine learning techniques to predict various sub topics within a text. Previous research (Clifton & Robert, 1999) shows it is possible to determent the subject of an entire text. However, research towards the identification of sub topics within a text is scarce. The methods this thesis will compare are Support Vector Machines, Decision Trees and Naive Bayes. It shows that there is no method that clearly outperforms the other methods however, on average, the Support Vector Machine does have the best average performance for prediction the sub topics within a text. The dataset which I use is gathered from various review sites.

Table of contents

Introduction

Before buying a product most people will try to gather experience from other people about that particular product or service (Pang & Lee, 2008). Initially people used to ask friends, family or a specialist about their experience with that product. These reviews were mostly seen as very reliable, because you would ask people whom you personally know. However, when asking such a narrow circle, chances are that nobody has experience with a certain product or service. This completely changed with the rise of the internet. Nowadays it is much easier to find reviews of products or services online. Although internet offers a huge amount of reviews on a certain subject, the value of an online review is not the same as from a 'real' person. However, it has been shown (Tirunillai & Tellis, 2014) that multiple online reviews are seen as reliable as a personal review of a product. This phenomenon is called "wisdom of the crowds". This (Tirunillai & Tellis, 2014) research also shows that 'wisdom of the crowd' does not apply to aggregated scores of a product; people will read entire reviews when trying to gather experiences. As shown in the next chapters, these reviews can have a big impact on the sales of a product. Therefore, a good review can have a big impact on product performance.

Previous papers (Pang & Lee, Opinion mining and sentiment analysis, 2008; Tirunillai & Tellis, 2014) have been based on complete reviews. In this thesis I will be looking to identify the sentiment of different subtopics of a review. I hope to find out which different subtopics are important for a good aggregate score. Knowing this, product designers can find which subjects are important for good sales, which can help designing new products.

Currently more and more systems use topic identification methods to mine vast amounts of texts. This is used for all sorts of purposes such as search engines, or topic-specific research. There are publications about topic identification for entire texts (Clifton & Robert, 1999). However, there has not been a lot of work done identifying different topics within a text. In this thesis I will compare different possible methods for identifying topics within game reviews. After identifying topics, I will predict the scores about the specified topic.

Being able to identify subtopics within text might lead to a better understanding of why certain groups buy different types of games. This could also help focusing on specific subjects when developing new products.

1. Problem statement and research objective

Using topic classification is a great way to get a sample of articles with the same topic from a big list of articles. This allows a researcher to gather more articles than was possible without the help of topic categorisation. Using commonly used sentiment analysis (Pang & Lee, Opinion mining and sentiment analysis, 2008) shows that it is possible to capture the defining sentiment for an entire article. The methods they review are mostly lexicon based. For example, they count occurrences or the presence of words to identify the topic of a text. However, it will be harder to identify the sentiment about specific subjects within an article. This is the aim of this research; to develop an easy to use method to give a sentiment about different subjects within an article. The research objective is therefore to find a good method for sentiment analyses about a specific sub topic within a text.

This gives me the following research question:

Which statistical method, discussed in this thesis, can be used best to identify and evaluate the sentiment of subtopics within a text?

2. Scientific and managerial relevance

My research will look at current methods of mining sentiments of reviews and different sources of reviews. I will compare different methods to each other to see which has the best fit on my data. Future scientific research will benefit from this. Knowing for future marketing research which methods could be used for sub topic identification and what performance to expect will be a good starting point for specialised sub topic research.

Businesses managers will always try to make as high as possible profits. Increasing sales will help realising this goal. As stated in the introduction, good reviews and not only aggregate review scores can increase online sales when people read these good reviews. However, at the moment there is no easy way for business owners to know the sentiment of reviews in order to easily analyse the comments. Giving business owners a tool to analyse these reviews so these can be ordered will properly increase sales of the web store and therefore the profits. This research aims to provide a good and widely applicable method for analysing the sentiment of specific topics within reviews.

3. Structure of the thesis

In the first part of this thesis I will look at the already existing research on the subject of sentiment analysis. Here I will list the methods used for this subject at the moment and make a theoretical comparison between these methods. In addition to a theoretical comparison, I will look for the effectiveness of prediction of the various methods. Here I hope to find out what is the most effective method for text analysis from previous research.

In the second part I will briefly analyse the current market for video games. Here I will also discuss the data sources that I will use in this research and the method used for gathering this data.

Thirdly, I will use the data that I gathered before to train the model for the prediction of the sub topics. The models I will use here will be explained in the theoretical part of this thesis. I will look at an initial performance of the sub topic identification per model. As the second step in my process, I will train models to predict the sub topic score. Also, here I will look at the initial performance of these models. Here I hope to see some initial results on the best method of prediction the sub topic of a paragraph.

In the fourth part I will apply all the models which I have trained in the previous part. Most importantly, I will look at the performance per method to predict the sub topic score. I hope to conclude here which method has the best predictive value of predicting the sub topic per paragraph.

Lastly I will look at the implication of my conclusions on managerial decisions and the relevance for future research. I will also look at the shortcomings of this research and the direction for future research.

Theory and hypothesis

1. Literature review

In the last years there has been quite some research on the subject of sentiment analysis of texts. This research is mostly done by computer science and linguistic departments. The result of this text analysis research is being used more and more in marketing departments, of which some examples are shown below. The actual problem of deriving the sentiment of a given text is not the main issue of the current marketing researchers. Therefore, most of the sources used are more technical publications instead of marketing papers.

Firstly, I will look at the research done by marketing departments to see the effect of reviews on sales. Afterwards I will look more closely to the techniques used to derive sentiments of reviews.

Paper	Industry	Estimation method	No of observations	Time frame	Main insights
Zhu & Zhang (2010)	Games	differences-in-differences	123	2003-2005	The findings indicate that online reviews are more influential for less popular games and games whose players have greater Internet experience.
Chevalier & Mayzlin (2006)	Books	differences-in-differences	3587	1998-200	Reviews do have effect on book sales, people read the entire review and extremely low scores have greater effect than extremely high scores.
Chrysanthos, Zhang, & Awad (2007)	Movies	Markov Chain Monte Carlo	80	2002	Reviews can help predict the future sales of a movie.
Ludwig, et al. (2013)	Books	Lagged linear regression	591 books and 18682 reviews	April 15 – May 5, 2010	The relation of positive reviews and negative reviews is asymmetrical on the conversion rate of books. Managers should find and promote the most influential reviews to stimulate online sales.
Senecal & Nantes (2004)	Online experiment	Generalized estimating equations	487	N/A	People who consult online reviews are more likely to buy a certain product then people who do not consult reviews.

Table 1 review of previous research focusing on the effect of reviews on sales

Table 1 shows a small summary of the previous papers about the effect of reviews on the sales of products. The overall outcome of the research shows that reviews have a significant effect on sales. To isolate the effect of reviews, Zhu & Zhang (2010) use a 'difference in difference' method. They compare the sales of the same game on the Xbox and the PlayStation 2. They assume all the other variables are the same on both the Xbox and the PlayStation 2 and only the reviews differ. Using this method, the sole effect of reviews can be measured on the sales of the games. Using sentiment analysis, the reviews are analysed. In this research this is done by counting the occurrences of words. Using this method, it shows that more positive reviews result in more sales.

Senecal & Nantes (2004) use an online survey to analyse more precisely the effect of online reviews. This study shows that reviews can steer people to buy a certain product. When people read reviews they are more likely to buy a product than before reading the review. This study also shows that different types of reviews have a different effect on consumer choses. Lastly, according to the research, personalised recommendations have the most effect on consumer choice.

Not only the overall sentiment is important for a review. When people can better identify themselves with a certain writing style in a review this will also add trustworthiness to a review (Ludwig, et al., 2013). This study also shows that negative and positive sentiments are asymmetrically correlated on conversion rates. The study also shows that the overall rating of a product has far less effect on conversation than real text reviews. To get the sentiment of a review, the linguistic inquiry and word count[1] programme was used. This programme gives a sentiment score to a text by counting positive / negative sentiment words. Since often the title of a review has a lot of meaning, the title is given extra weight in the overall sentiment of the text.

Above, the effect of reviews on the sales of products was briefly discussed. Unfortunately, there has not been research towards the effect of the sub topics on games sales. This sales effect of sub topics on sales will be outside the scope of this thesis. However, there has been research on the effect of product features on sales, below there is a short summary of the previous research done in this area.

[1] http://www.liwc.net/

Research towards novel features on products (Mukherjee & Hoyer, 2001) shows that a new feature can increase the sales of a product. This holds if the product is not too complex, when the product is more complex a novel feature can even lead to lower sales. This effect is studied in more detail (Thompson, Hamilton, & Rust, 2005) and shows that consumers initially choose products with a lot of features which actually results in a lower life time value. Therefore, this research recommends to build more specialised products, each with a special set of features instead of one product with all the features. Another research (Archak, Ghose, & Ipeirotis, 2011), which is similar to my research, also shows that product features have a significant effect on product sales. This research uses the unsupervised machine learning method to identify topics that are discussed in a large number of reviews.

Park, Milberg, & Lawson (1991) show that product features do not always have a positive effect on the sales of a product. Here it shows that it is important for product features to align with the brand expectations. If this does not hold, additional product features could negatively influence the product sales. Other research (Simonson, Carmon, & O'curry, 1994) also suggest that adding more product features does not always increase sales. Here it shows that is an additional product feature which is very small this could even lead to a decrease in sales.

These previous papers show that product features can have a big effect on the sales of a product. This research hopes to identify specific sub topics within reviews and the score of these sub topics.

Sub topic identification methods

Next I will look at various ways of deriving a sentiment of a review. A quick summary can be seen in Table 2 the methods will be discussed in more details below .

Method	Summary
Support Vector Machine (SVM)	Classifier, uses a hyper plane to separate the data set into two different groups.
Naive Bayes	Uses prior and evidence to calculate the probability a new data point belongs to a group.
Decision Tree	Builds a tree with certain decisions which would lead to a group classification
Neural Network	Uses multiple layers of nodes to process data step by step

Table 2 Summary of used methods

Each of the methods I will use are supervised learning methods. Generally, there are three types of learning methods, namely supervised, semi-supervised and unsupervised (Zhu X. , 2005; Yarowsky, 1995). Supervised methods have 'answers' the model tries to find a relationship between the dependent and independent variables, similar to a normal regression model. While unsupervised methods do not have 'answers', they try to model data only with dependent variables (Yarowsky, 1995). Semi-supervised methods are a combination (Zhu X. , 2005); they first learn with 'answers', like supervised methods. Then it will predict new data points, when the algorithm doubts the answers it will ask the researcher the proper answer. A more schematic overview can be found below in Table 3. In this thesis I will only make use of supervised methods. Previous research (Popescu & Etzioni, 2007) has used unsupervised learning methods to extract the product features form a large number of reviews.

Method	Supervised	Semi-Supervised	Non Supervised
Dependent variable	Outcome variable	Partly Outcome variable, partly none	none
Independent variable	Input variables	Input variables	Input variables
Examples for use	Regression / classification	Regression/ classification	Clustering

Table 3 Machine learning methods

Previous research (Taboada, Brooke, Tofiloski, Volt, & Stede, 2011) states that there are two main approaches of text mining, namely the lexicon-based approach and the text classification approach.

Within the text classification approach there are various methods used (Pang, Lee, & Vaithyanathan, 2002). The **Naive Bayes method** is a simple, widely used model for text classification. Although this model is not the most advanced, the research above shows that it is a very good predictor of sentiment. The Naive Bayes method calculates the probability of a certain event given data about that event. For example, when there is a group of red and green balls, when a new ball enters the dataset the Naive Bayes method gives a probability whether this new ball is red or green, given the data known about the ball. This example will be discussed in more detail below.

The Naive Bayes is formulated as follows (Zhang, 2004):

$p(C_k|x_1, \dots \dots, x_n)$

This gives the probability that the independent variables, represented by the vector $x_1, \dots \dots, x_n$, belongs to the class C_k. This can be rewritten to a more readable version (Smola & Vishwanathan, 2008):

$$Posterior = \frac{prior \times likelihood}{evidence}$$

This can easily be explained using a classification of two classes, as can be seen in Figure 1 (Dell, n.d.). The chance that a new data point belongs to a certain group is depending on the initial prior and likelihood. The priors are calculated by $\frac{Number\ of\ x\ objects}{number\ of\ y\ objects}$. Next, the likelihood is depending on the observations near to the new data point as can be seen in Figure 1. It is calculated by, $\frac{Number\ of\ x\ objecs\ near}{number\ of\ y\ objects\ near.}$. When these two outcomes are calculated you get the probability of the new data point belonging to either x or y.

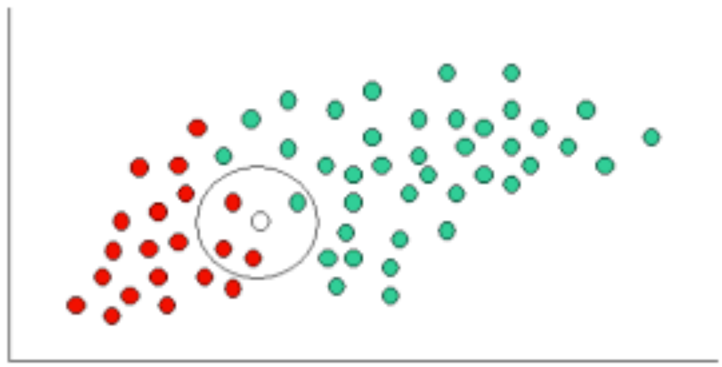


Figure 1 Naive Bayes classification source: dell.com

The same method could be used in text mining. For example, when there are two types of texts divided by having negative or positive sentiments. When all text with a negative sentiment contain a lot of negative words and a new text also has a lot of negative words, the probability that the new text is negative increases. In this thesis I use this method as described above. First, I will train the algorithm and then apply it on new data points. It is also possible to use this method as an unsupervised method (Dougherty, Kohavi, & Sahami, 1995), however I will not do so here. Although the algorithm assumes independence between the variables, which does not strongly hold in text, the algorithm performs well (Joachims, 1998).

The Bayes method uses probabilities to classify a review to a certain group, the **Support Vector Machines** (SVM) does not use probabilities (Smola & Vishwanathan, 2008). This method also classifies as a supervised learning method. It needs data with an 'answer' to train. This is similar to the Naïve Bayes method. This method can be applied well to binominal distribution, like the positive-negative sentiment analysis or sub topic analysis. This method tries to search for a hyperplane (see Figure 2) as large as possible between the two groups in the data set. The hyperplane describes the plane between the two data sets. The dotted lines in Figure 2 represent a hyperplane. When a new data point is entered, the hyperplane decides the class of the new data point (Friedman, Hastie, & Tibshirani, 2001). This hyperplane can be easily seen in a graphical illustration:

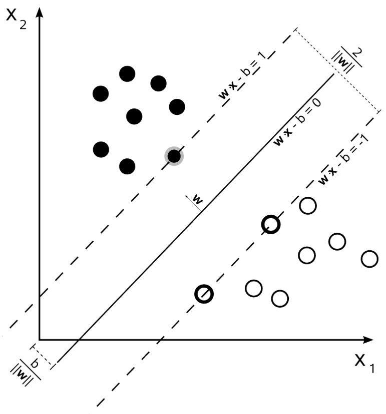


Figure 2 Support vector machine illustration source: Wikipedia.org

As can be seen there are two groups (black and white dots), which are separated by a hyperplane. When a new review is processed, the method looks on which side of the hyperplane it belongs to categorize it as a positive or negative review. This method outperforms the simpler Naive Bayes method in various researches (Joachims, 1998; Annette & Grzegorz, 2008).

Another commonly used method for machine learning is **decision trees** (Quinlan, 1986). This method has also been used in previous works (Annette & Grzegorz, 2008) for sentiment analysis, which showed that the SVM had a better performance. This method builds a tree with decisions on different points, as is shown in Figure 3. This models the change of surviving the Titanic disaster. As can be seen, the initial decision is about whether the person is a male or female. If the person is a female she would have survived the Titanic disaster. If the subject is a male and older than 9.5 years old he would not have survived. Lastly, if the male is older than 9.5 years and has more than 2.5 siblings on board he would not have survived either. In this thesis I will use this method to predict the subtopic of a certain paragraph.

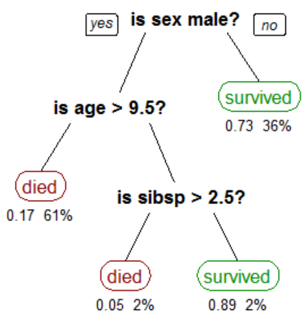


Figure 3 Titanic survivors decision tree source: wikipedia.org

Sub topic score prediction method

I will use the previously discussed models to determine the sub topic within a text. I have chosen these methods based on the previous researches, which are mentioned above. I will now introduce another model, namely neural networks. I will use this method to predict the score of a sub topic.

Lastly, I will use **neural networks** for the sentiment prediction. Like the systems above, I will train the neural with sample data before new data point can be assessed. This is not necessary as it can also be used as an unsupervised learning method (Bishop, Neural networks for pattern recognition, 1995; Sanger, 1989). Neural networks are composed of different neurals. This can graphically be represented as follows:

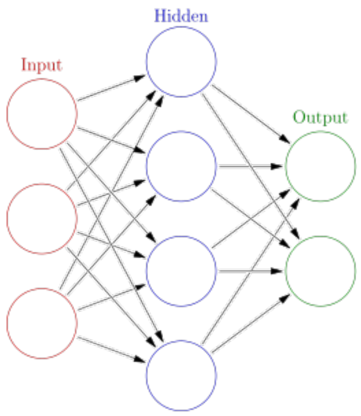


Figure 4 Neural network | By Glosser.ca [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons

Figure 4 represents an artificial neural network with only one hidden layer. As can be seen there is only one hidden layer of neurals between the input and output layers. The hidden layers are the layers between the input and output layers. It is also possible to make neural networks with multiple hidden layers. An example can be seen in Figure 5. This is an example with the same data from the Titanic as used above with the decision tree.

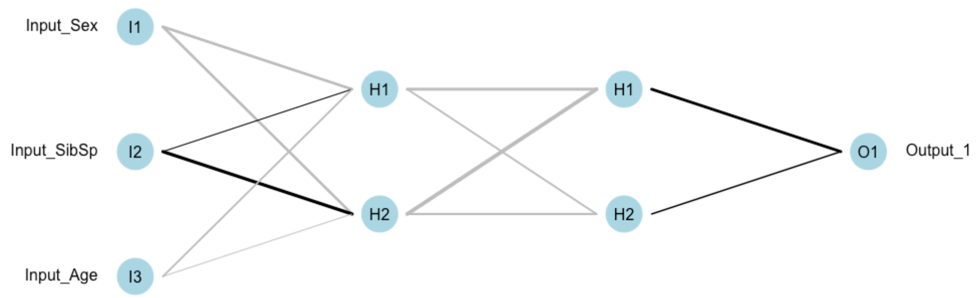


Figure 5 Multi-layer Neural Network Titanic Data

The input from the independent variables will go into the first layer this is called the input layer, the input will be partly processed in this first layer. The input layer will pass the partly processed data to multiple nodes in the hidden layer for further processing. After the processing of the hidden layer(s) the data is passed to multiple nodes in the output layer. This method has been used before for sentiment analysis of Twitter messages with good results (Ghiassi, Skinner, & Zimbra, 2013).

I will use the models which I initially proposed (SVM, Naive Bayes and Decision Tree) to predict the sub topic of the paragraphs. Next, I will use the neural network to predict the scores per sub topic. The methods will be discussed in more detail below in the conceptual framework.

2. Conceptual framework

The goal of the thesis is to predict the sub topics within a text. Having this information, I hope to be able to quantify the meaning of a review. This could be used by product designers to improve product design or for researchers to find relationships between the sub topics and sales.

Because I want to predict the sentiment of a certain sub topic of a text I will first have to predict the subject of a certain paragraph of the text. I define a paragraph as two lines of text separated by one full text-less line. I will compare three different methods of classifying the subject of the paragraph. I will look at three sub topics: Graphics, Gameplay and Audio. I have chosen these topics because most online review sites with subtopic scores use these subtopics. Figure 6 shows a graphical representation of the methods which I will use. This will be discussed in more detail below.

Determine the sub topics within a review

The first method I will use is a simple counting method to determine the sub topic per paragraph in the review. I will count in which paragraph the words correlated to a sub topic are discussed the most. For this I will use a wordlist with words corresponding to sub topics. For example, when I want to identify the subtopic 'graphics', I will use a wordlist such as "graphic, graphics, looks, …". To create these lists I will manually analyse reviews and select relevant words for a specific sub topic. I have tested these wordlists by comparing the words that I have found and the most used words which I have found in the manually coded graphics paragraphs (see below for more information about the manual coding). I will assume that the paragraph with the most matching word will be about the graphics of the game.

The other methods I use make use of the different machine learning methods which I have discussed the in the theory part of this thesis. I will manually code a subset of the reviews and classify the subtopic of each paragraph. I will then use this data to train a Support vector machine, Naïve Bayes and Decision Tree to identify a certain subtopic. When the models are trained, these models will predict the subjects of the paragraphs which I did not classify manually. The paragraphs which were not defined with a sub topics will be removed for the next step.

Determine the score for each sub topic

When I have the certain subtopic of a paragraph I will train a neural network to predict the subject sentiment from the paragraph text. I have chosen this method because the performance in previous research (Ghiassi, Skinner, & Zimbra, 2013) using neural networks has proven to be very successful. For testing in the next phase I will use the derived neural network to predict the sub topic score of paragraph which has not yet been classified.

Determine the model performance

Finally, to test which method has the best predictive value, I will look at the difference between the predicted scores for each method and the scores form the review. I will assume that the neural network will always perform the same, therefore the difference between the goodness of fit will be because of the different classifying methods. I will use different metrics for comparing the actual score with the predicted score. I will look at different metrics because the metrics I use all have a different emphasis. The used metrics are the Root Mean Square Error, Mean Absolute Error and Root Mean Squared Logarithmic Error, these will all be discussed in the model performance part at the end of this thesis.

This way I use the difference between the actual sub topic score and the predicted sub topic score as a measure.

This outline of this method can be shown below in Figure 6.

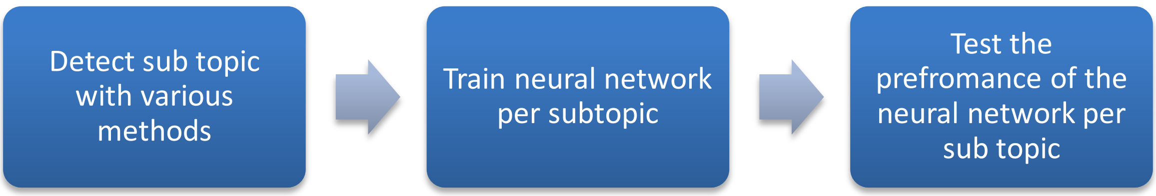


Figure 6 Conceptual framework

3. Dependent variable

Because my research design has multiple steps, I do not have one dependent variable. Although the most important dependent variable will be the sub topic. Because the main goal to create a model which is capable of detecting the sub topic of a paragraph.

Thus, in the initial step the sub topic will be the dependent variable. This variable is a Boolean variable; the paragraph can either be or not be about a certain sub topic. Because this variable is Boolean the data does not need to be transformed to be used with the proposed models (SVM and Decision tree).

Secondly, I use the sub topic score to predict my Neural Network. So this will be the second dependent variable. This variable is a continuous variable that takes values between 1 and 10. Neural networks are capable of using this as a dependent variable, so I do not need to transform my data to use this method.

4. Key Independent Variable(s) and their Relation with the Dependent Variable(s)

As independent variables I have used the review texts. However, I am not interested in the complete review text since I want to know the sub topics within a text. Therefore, I will transform the data before using it in my models.

Initially I will split the text in paragraphs so that I can determine the sub topic per paragraph. Because it will be hard to analyse a text, I will transform these texts into a Document Term Matrix (Li & He, 2009). The method to do this will be discussed in the next chapters.

Data

1. Empirical setting

This thesis is about the gaming industry, which yielded a total sales of 15 billion dollars in the US in 2014. This is excluding the sales from game related articles like consoles (Entertainment Software Association, 2014). The total gaming business is calculated at 22.41 billion US dollars in the US in 2014. This includes games and accessories such as consoles. The previous report shows that the industry has been growing fast from 2003 until 2010 when sales stagnated. Looking at the worldwide market and market trend, Fortune (Gaudiosi, 2015) predicts that the total gaming market in 2015 will be worth 91.95 billion US dollars. This article predicts that mobile games will have more sales in 2015 than PC or console games, which would be the first year ever that mobile games have more revenue. The Newzoo research company predicts (Newzoo, 2015) that the gaming market will show an average annual growth of about 7.9 %. This shows how important the gaming industry is becoming in the world economy.

The data which I will use here is collected from gaming sites, which aim at providing gamers with useful reviews. Using this source, I hope to get data which is the most applicable in real life. As seen in the previous chapters, people use online reviews to base their purchase decisions on. Therefore, the data that I am using is indeed applicable to real life decision making.

2. Data collection

To collect my data, I used a web-crawler, specifically 'Import.io'. A web crawler is a piece of software which saves information from publicly accessible websites. In this case the web crawler will open web pages with gaming reviews and save the reviews and scores on that page. This crawler allowed me to quickly and accurately crawl my specified sites. I have used the following to collect data:

1. GamrReview http://www.gamrreview.com/
2. Gaming Target http://www.gamingtarget.com/

I use these sites because they not only give a score for the entire review but also a rating per sub topic. Moreover, these sites offer reviews which are very structured. This makes it easier to analyse the text. Lastly, these sites are actively maintained and have a very large number of reviews in their database. This allows me to have a large training set which will be very useful in the next chapters. In the Appendix Picture 1 a sample page of Gaming Target can be found. The crawler looks for the following elements:

- Title
- Total score
- Visual score
- Audio score
- Gameplay score
- Review text

The two websites which I use for this research are only a very small portion of the total number of websites which offer gaming reviews. However not all review sites offer sub topic scores which is necessary for this research.

3. Definition of measures

The data I will use is mostly text data. Because this data is not easily processed I will transform these texts to a Document Term Matrix (DTM) this process is also used is previous research (Dhillon, 2001). The DTM is a matrix with all the possible words in all the texts as columns and the occurrences per paragraph in the rows. To reduce the noise in the data I perform different computations before creating the DTM. These steps will be necessary for further analysis of the text. If I do not do this the number of variables will be too high and the variables might be irrelevant. Moreover, previous research (Dhillon, 2001) shows that using this DTM increase the possibility to predict scores. I will use these steps:

1. First I make all words lower case – If I would not do this the words ‘Good’ and ‘good’ would be different variables and therefore I transform everything to lower case.
2. Remove all punctuation – Because I look at word occurrences the punctuation does not have a function. Moreover using the same example as above I want ‘good,’ to be the same as ‘good’.
3. Remove all numbers – I am looking for words within text and therefore I remove all numbers.

4. Remove stop words – Languages have a lot of stop words like ‘the’, ‘and’ and ‘it’, previous research shows these do not give extra predictive value to the DTM (Dhillon, 2001).
5. Stem the document – This turns every word to its root. For example, ‘reading’ and ‘reads’ will be stemmed to ‘read’. The algorithm I use here is prevalent, this searches for the most frequent stem of the word.

Before I start training the model I will remove words which do not occur more than 4 times in the entire training set. I hope to remove spelling and other mistakes this way. Previous papers (Ghiassi, Skinner, & Zimbra, 2013) show that this method of removing such words does not decrease the predictive value of the models.

I think using this method removes most of the noise in the data set and it will be feasible to apply the computer-intensive methods like neural networks.

In addition to the text I will use other variables which are shown in Table 4. As can be seen the other variables are nominal variables from 1 to 10.

Variable name	Description	Measurement	Type
Total score	The total score which is given the a certain game	0.00 to 10.00	Nominal
Gameplay	Gameplay score of review	0.00 to 10.00	Nominal
Audio	Audio score of the review	0.00 to 10.00	Nominal
Graphics	Graphics score of review	0.00 to 10.00	Nominal
Review	The review it self	Text	Text

Table 4 Definition of measures

4. Data description

Before doing further analysis it is important to look at the data. Initial summary statistics can be found in the appendix A Table 11 . The GamrReview has almost 1000 reviews while the Gaming target set has more than 2000 different reviews. As shown, both the data sets initially look quite the same. For example, the average scores and deviations for the variables are highly similar.

Table 12 and Table 13 in Appendix A show the correlations between the different variables of both the data sets. As seen in both datasets the characteristic ‘gameplay’ has the highest correlation with the total score. This can indicate that the gameplay can have a big influence on the total game score. This will also be discussed later.

Analytics and Results

First I will further analyse the effect of various characteristics of the game on the total score. Afterwards I will look at the analysis of the results of the reviews.

1. Effect of features on total game score

Before looking at different methods to get the subtopics of the reviews I want to see if there is a significant effect of feature scores (gameplay, visuals, audio) on the total score of the game. I will do so by using a simple linear regression model (Moore, McCabe, Duckworth, & Alwan, 2009), with the total score as dependent variable and the features as independent variables. I have chosen for this simple model because it is easy to use and the data roughly fits a linear relation as shown in Graph 8. I will include the author of the text as a fixed effect. Because I wonder whether different authors have different methods of judging various subtopics and overall score. The entire function is shown at Equation 1 the $F_{1,..,n}$ denotes the features per game and the author is the fixed effect.

$$overalscore = \beta_0 + \beta_1 F_1 + .. + \beta_n F_n + \gamma_0 author + \epsilon$$

Equation 1

Model	Intercept	Gameplay	Value	Presentation	Adjusted R^2
1	1.136***	0.844***			0.818
2	2.420***		0.700***		0.739
3	2.107***			0.718***	0.627
4	0.020	0.412***	0.292***	0.298***	0.969

Table 5 Regression output game review

Model	Intercept	Gameplay	Visuals	Features	Audio	replay	Adjusted R^2
1	1.453***	0.807***					0.820
2	1.869***		0.754***				0.563
3	2.155***			0.738***			0.695
4	2.696***				0.662***		0.526
5	3.049***					0.645***	0.746
6	0.145**	0.355***	0.171***	0.153***	0.135***	0.176***	0.945

Table 6 Regressions output Gaming target

*** P value below 0.01
** P value below 0.05

The results are shown in Table 5 and Table 6. I have not included the fixed effect because this effect did not change the coefficients of the model and the effect was not significant. I use all variable available per data set. Unfortunately not both datasets have exactly the same variables. As can be seen in each model all the characteristics of the game have a significant effect on the total score. However not all models are able to have a good predictive value. For example, Gaming target model 2 only has an Adjusted R^2 of 0,563. Both datasets show that gameplay has significant and large effect on the total score with a high Adjusted R^2. Looking at the models with all the characteristics the gameplay still has a large effect, especially in the GamrReview dataset.

This is very important for the value of this thesis. This shows that sub topics have a significant predictive value for the total score of the review. Knowing this, a total score can be predicted form the sub topic scores which I will predict in the following chapters.

2. Defining the subject of a paragraph

To be able to get the subject of a paragraph, I have set up three different methods. Two methods use machine learning while one method does not need any additional training before it can be used. This is the first step in the two-step process to predict the sentiment about a specific subject. As statistical software I will use R for all computations (R Core Team, 2015). The data set which I will use is Gaming Target because this set has the best structured data. I did include the GamrReview data in the previous part to see if the sub topics have the same effect on the total score like in the Gaming Target data set. Because the sub topic effect is the same, it is more likely that the relation between the sub topic score and total score exists as well in data sets which I have not researched. The sub topics which I will investigate are ‘gameplay’, ‘audio’ and ‘visual’. I have only used these because when looking at different review sites it shows that almost all sites have these three sub topics.

Figure 7 shows how the data get processed for the machine learning methods. The entire process will be discussed in more detail at the Support Vector Machine topic.

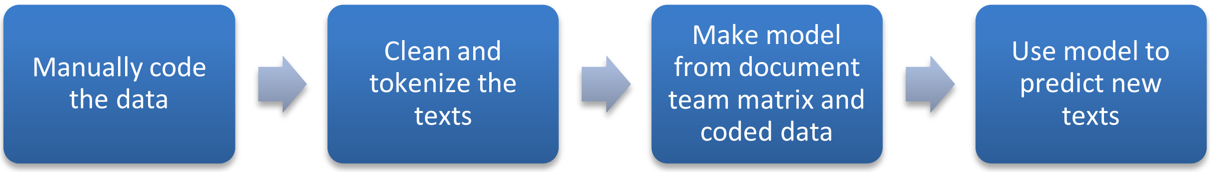


Figure 7 Process for sub topic identification

The methods I will use are the following:

1. Simple counting of words.

   This method looks where in a text most words associated with a sub topic are counted. Therefore, this method requires a word list associated with the subject you want to find. Here I use wordlists with words associated with 'gameplay', 'audio' and 'visual'. I have manually coded these wordlists after reading almost 2000 paragraphs. Next, I count the number of occurrences per subject per paragraph. I then select the paragraph with the most occurrences as the paragraph about the subject. A consequence of this method is that all sub topics will be identified in every text, even though the review might not speak about that sub topic. This method can be compared like the method used by Pang & Lee (2008).

2. Using a Support Vector Machine

   Here I make use of a SVM, as described in chapter two. To make use of the SVM first a model has to be trained. To do so, I have manually coded 1600 paragraphs, for every paragraph I coded whether it was about 'gameplay', 'audio', 'graphics' or neither. This model cannot be trained using the plain text of each review. Therefore, I make a Document Term Matrix, as discussed before.

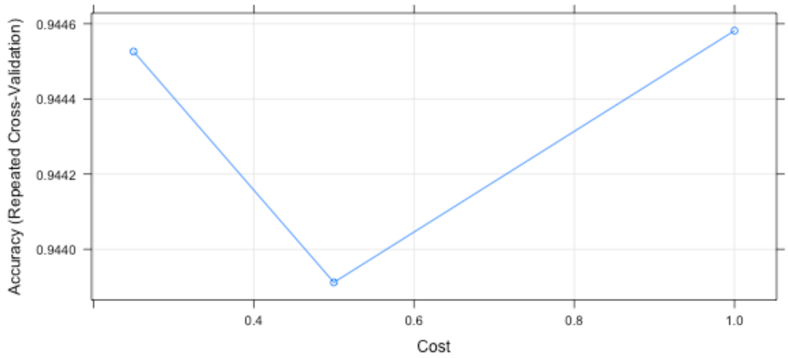
   I will train the model SVM model in R (Dimitriadou, Hornik, Meyer, & Weingessel, 2008). The model uses multiple parameters that affect the model (Karatzoglou, Meyer, & Hornik, 2005). Two important parameters are Gamma and Cost (C). The gamma parameter defines the influence of one training input, in this case one paragraph. The Cost tells the model how precisely to fit the hyper plane a higher C gives more precision, but this can lead to over fitting. Over fitting (Vapnik & Vapnik, 1998) can occur when the number of variables is very large but the number of observations is small. This will decrease the ability of the model to predict new data points.

After the model is trained I applied the model to the paragraphs which I did not use for training. The outcome shows the predicted sub topic per paragraph. The disadvantage of this method is that not all reviews will have a paragraph about every subject. Therefore, for some texts not all sentiments can be predicted. This could also mean that within one text multiple paragraphs could be selected to discuss a certain subject. However, I believe this could increase the fit of the neural network in the next step.

The training results yield the following which can we seen in Table 7 and Graph 1. Table 7 and Graph 1 shows various models that are trained with different values of C and Gamma (Kuhn, 2015). The value for Gamma has been constant at the level of 0. This means that the SVM does not have a better performance with other values for Gamma. The best model is chosen which has option Cost =1.0 and Gamma= 0. This model is chosen because it has the highest accuracy. Accuracy is calculated by training a model and then applying it to the training data set. The model with the most accurate prediction is chosen. When the best model is chosen, this will be used in the prediction for the future calculations. Graph 1 shows a plot of the different values for parameter Cost and the Accuracy, also here it shows that the Accuracy is highest when cost is 1.0.

cost	Accuracy	Kappa	Accuracy SD	Kappa SD
0.25	0.9445260	0.4714603	0.007462960	0.07302714
0.50	0.9439119	0.4667668	0.007627367	0.07201202
1.00	0.9445816	0.4653690	0.006975878	0.07644847

Table 7 Partial outcome svm graph



Graph 1 Accuracy for different values of Cost

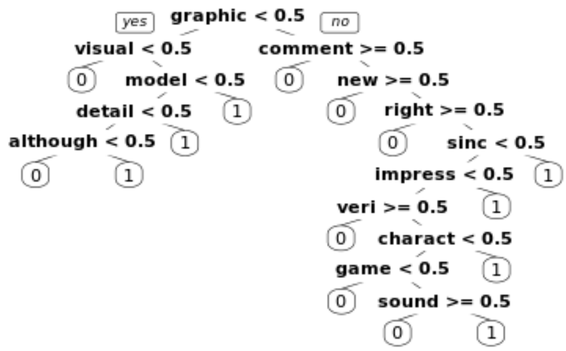
3. Using Naive Bayes

    Thirdly, I used the Naive Bayes classification. The method is the same as for the Support vector machine. This method will also have the same downsides as the SVM methods. Some texts will not have paragraphs about a certain subject. This has advantages and disadvantages, as discussed in the previous part.
    The big difference however is a Naive Bayes classifier gives a probability, as explained in chapter two. Same as with the SVM I will use the Caret package in R to train and predict the model.
    Unfortunately, the results obtained with this method were not significant and did not show to be a good method for predicting the sub topic. The outcome showed that this method never classified a paragraph as a sub topic. Therefore, I do not take this method into consideration.

4. Using Decision Tree

    Lastly, I used Decision Tree to predict the sub topic of a paragraph (Quinlan, 1986). Here I will also use the same method as with the SVM. The main parameter which is used is the depth of the tree. This means the number of levels in the tree, Figure 3 shows a tree with 3 levels. A part of the final tree for the sub topic graphics can be seen in Graph 2. Here outcome 1 means the sub topic is about graphics. The outcome of the tree which is used to predict the subtopic Graphics is shown in Table 8 and

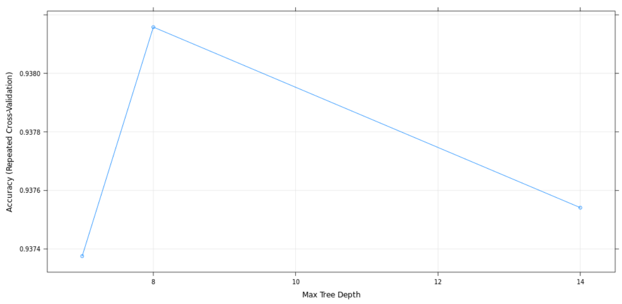
Graph 3. Table 8 shows the different levels of the maximum levels in the tree and the accuracy per maximal level. The goal is to have the highest possible accuracy, therefore the maximum depth of 8 is chosen (Kuhn, 2015). In Graph 3 the same data is shown in a graph, here is also shown that the highest accuracy is with the max levels of 8.



Graph 2 Decision Tree

Max depth	Accuracy	Kappa	Accuracy SD	Kappa SD
7	0.937	0.372	0.012	0.135
8	0.938	0.380	0.012	0.142
14	0.937	0.388	0.013	0.138

Table 8 Outcome Decision Tree Graph



Graph 3 Outcome Decision Tree graph

3. Predicting the score of a paragraph

As the second and last step I will try to predict the score of a review about a certain subject. In the previous part I determined the subject of each paragraph. Each of the scores is a nominal variable between 1 and 10. Previous research (Pang & Lee, Opinion mining and sentiment analysis, 2008) has shown that neural networks have a good performance with predicting the sentiment of text and therefore I will use Neural Networks here to for the score prediction. I will do so in the same manner as training the previous models. To train the model I will first predict the subject of 5000 paragraphs using the different methods described above. Of these 5000 paragraphs only a sub set will have the subject which was chosen. For example, when using the SVM method for graphics only 134 paragraphs where selected as having a graphics topic so only these will be used for the training. With this subset the neural network will be trained. When the model is trained I will use this model to predict subject scores for other paragraphs. To test the performance of these models I will compare the predicted results from the neural network with the actual subtopic scores values. I will presume the neural Network will always perform the same for each of the subject prediction methods. Therefore, I presume the model which performs the best is due to the subject prediction. This method can be shown graphically as in Figure 8.

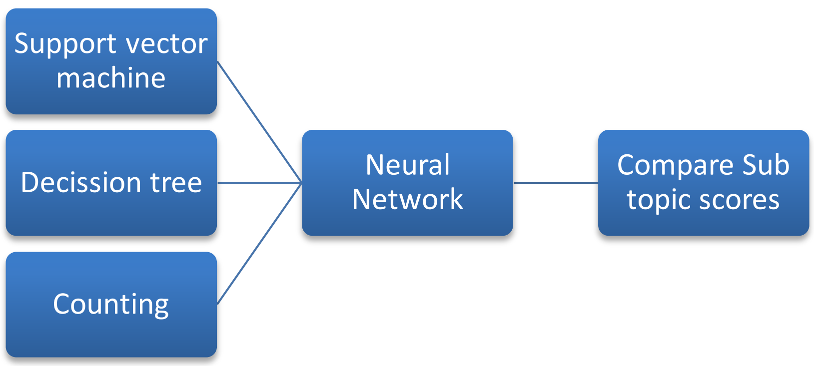


Figure 8 Complete process

Again, same as in the previous part, I have used various different parameters to tune the models. To start the tuning parameters, I specify the number of nodes in hidden layers. The maximum number of hidden layers is three. The options are 100, 50 and 30. I have chosen these values after experimenting with different values and it showed that these give the best results. I have tested with values ranging from 0 to 100. Besides the number of nodes in the network, I have also tried various values for the learning rate of the network. These values mostly lie between 0.1 and 1. Lastly, before training the model I removed words which are used too rarely; I remove all words which are not used more than 7 times in 5000 paragraphs (Ghiassi, Skinner, &

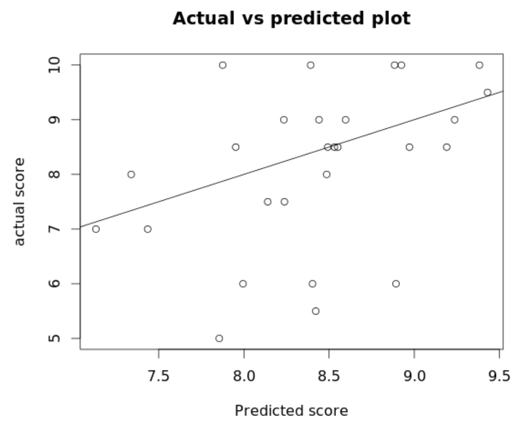
Zimbra, 2013). Like with the SVM I tried to rule out typing mistakes and words which are extremely rare.

To make a fair comparison between the different models, for subject identification I will train the neural network with the equal number of observations. In the case of Audio all the methods will only have 79 training paragraphs, because the SVM method only identified 79 paragraphs with this sub topic, although other methods did identify more paragraphs. This makes the comparison between the models more equal. Because increasing the number of observations in this case paragraphs increases the fit of the model. Therefore, if I would do this for one sub topic the comparison between the different methods would not be fair.

The model with the Lowest Root Mean Square deviation is chosen. The RSME is the standard deviation between the sample and the predicted data. Therefore, the model with the Lowest RSME value has the best fit in the sample.

Below the results of the neural network with sub topic graphics topic and the SVM classification are shown. This gives an example on how the neural network preforms. The final model was tuned with 3 layers, having 100, 30 and 50 nodes in each respected layer. This result is given from the statistical program, more technical details can be found in the appendix B. Below Graph 4 a scatter plot between the predicted scores and actual scores is shown. If the model would be perfect this scatter would show a perfect 45-degree line. In the graph a 45-degree line is fitted to easily see the error. As can be seen the predicted values do roughly follow the line, however there are outliers. This means the neural network is not able to correctly predict these outliers.

Secondly, in Table 9 the performance measures of the final model can be seen. Because the model has been trained and the best model is chosen, there is not a single RMSE and R squared value. An extra explanation on the RMSE value can be found in the next part. The RMSE metric is quite large, this could be explained because of the outliers. The RMSE metric gives more weight to bigger errors.



Graph 4 Actual vs predicted neural network

RMSE

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
fit	0.8342	1.276	1.476	1.551	1.713	2.731	0

R squared

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
fit	0.000	0.01584	0.0704	0.1109	0.1506	0.5330	0

Table 9 performance measures neural network

4. Model performance

When both the SVM, decision tree and neural network have been trained I will use these together to predict the various subject scores with the actual subject score. As stated earlier, these models will not predict all the scores. Because in quite some reviews the SVM and or Naïve Byes models does not find a paragraph with a certain topic. As prediction set I will use the data which I have manually coded. I use this data set instead of a new test data set because when I use the manually coded data I am sure this paragraph is about the sub topic which I am testing. This could not be guaranteed when I would use a data set that is not manually coded.

To test the predictive value of machine learning various measures can be used (Peng, Wu, & xu, 2011). These methods all look for the relation between predicted (fitted) value and the actual value. I have used three measures to test the performance per model; RMSE (Root Mean Square Error), MAE (mean absolute error) and RMSLE (Root Mean Squared Logarithmic Error) (Tianfeng & Draxler, 2014). These measures are calculated as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(y_i) - \log(\hat{y}_i))^2}$$

The RSME measure initially calculates the difference between the actual value and the predicted value. Secondly, the difference is squared. Thirdly, the mean of all these squared differences is calculated. Lastly, the root of this mean is taken. This method gives more weight to bigger differences. For example, an error of 2 is more than twice as bad than an error of 1.

MAE simply takes the mean of the absolute difference between the predicted and actual value. This means all differences are weighted equally, this is the mean difference the RSME method.

Finally, I use the RMSLE. This method initially takes the logs of both the predicted and fitted value. Secondly, the difference is squared. Thirdly, the mean of all these differences is taken. Lastly, the root of this mean is taken. This method gives less weight to bigger differences. For example, the value of RMSLE would be roughly equal for the difference between 100 and 10000, as for the difference between 1000 and 100000.

Beneath Table 10 show the RMSE, MAE and RMSLE results for all the sub topics. As can be seen, there are three different methods for predicting and the corresponding RMSE, MAE and RMSLE scores. If these scores are lower, it means the model is better able to predict the scores. As shown, the easy count method is the best able to predict the audio scores, because this method has the lowest RMSE and MAE scores. In

Graph 5 the actual VS predicted values are plotted for the audio subtopic. Also, the ‘perfect’ line is plotted. With a perfect prediction all the values should follow this line. As can be seen it is very hard for all the methods to predict the lower scores.

However, the results which are found with RMSE and MAE can be confirmed by the graphs, showing that the count method does follow the line better than the other methods.

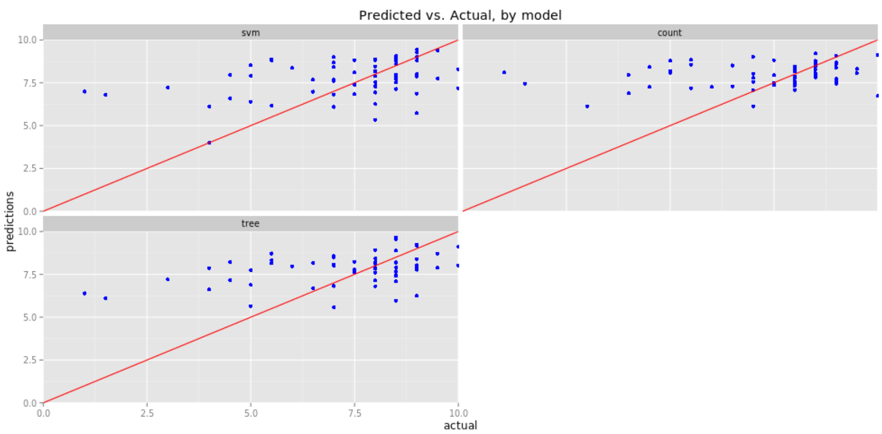
Therefore, it can be concluded that when audio is the subtopic, the best way to predict the sub topic is by using the Count method. Even though this is the case, it should be noted that this method is not able to predict the lower scores well, as can be seen in the graphs. Although the performance of the score prediction is not very good, this research is about the differences in the sub topic identification. Which is still possible with these results from the neural network.

Table 10 also shows the results for the other sub topics. As can be seen in the table there is no clear method which always has the highest predictive value. Also, when looking at the other actual vs fitted graphs namely Graph 7 and Graph 6, there is no clear method that performs better than another method.

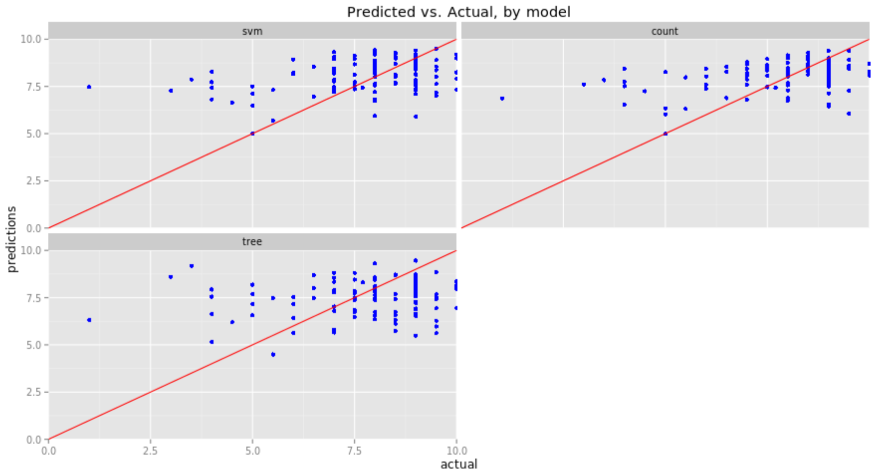
To conclude, to select best method for sub topic identification I look at the average performances of the different methods, which can also be found in Table 10. As can be seen on average the SVM methods has the lowest average scores. This means that the SVM models are on average the best to predict the sub topic of a certain text.

	Graph			Game			Audio			Average		
Method	RMSE	MAE	RMSLE	RMSE	MAE	RMSLE	RMSE	MAE	RMSLE	RMSE	MAE	RMSLE
SVM	1.65	1.24	0.24	1.78	1.33	0.25	1.88	1.36	0.31	1.77	1.31	0.27
COUNT	1.58	1.16	0.23	2.09	1.61	0.29	1.98	1.34	0.33	1.88	1.37	0.28
Tree	1.82	1.41	0.26	1.82	1.32	0.26	1.82	1.36	0.33	1.82	1.36	0.28

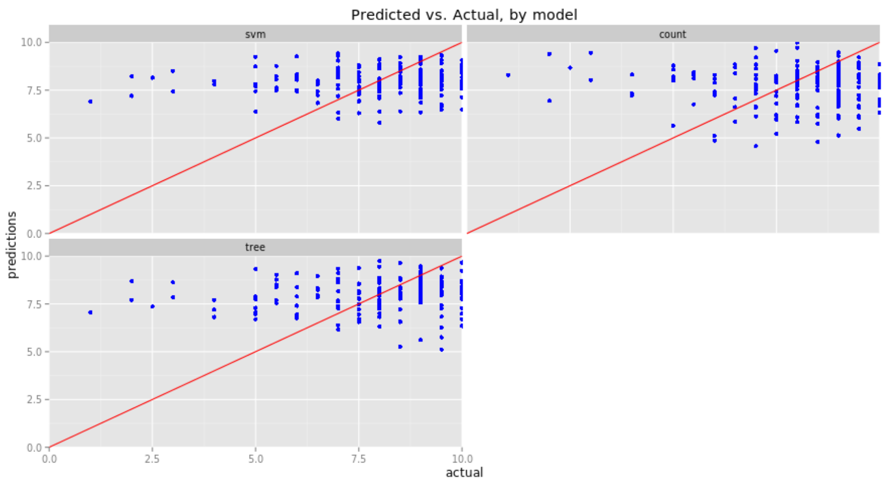
Table 10 RMSE and MAE results



Graph 5 Actual vs Fitted values audio



Graph 6 Actual vs Fitted Graph



Graph 7 Actual vs Fitted Game

Conclusion

1. General Discussion

This thesis has looked at different methods to predict sub topics within a text, and predict the score about this sub topic. I have used two supervised machine learning methods and one lexicon based method. This can be seen as an addition to the previous research which uses different cluster techniques to identify the topic of a text (Popescu & Etzioni, 2007). My results show that using a Support Vector Machine is the best method to predict the sub topic within a text. My research also shows that using the Naive Bayes machine learning method is not properly able to predict the sub topic within a text. Based on the previous research of Ghiassi, Skinner, & Zimbra (2013) I use Neural Networks for the prediction of the subtopic scores. My research did not get the same accuracy as this previous research, my methods were not able to properly predict the lower sub topic scores. Lastly, this thesis has shown that using different methods for sub topic identification does make a difference, which can be used in future research to search for even better methods to predict the sub topic of texts.

2. Academic Contribution

There has been a lot of research to topic identification on complete texts for example Popescu & Etzioni (2007). My research did not find a lot of previous research about sub topics within a text. There is research towards clustering text which also splits texts into sub topics. However, with clustering it could be hard to get the exact topic of a cluster, because most cluster methods are unsupervised learning methods. This research could be a starting point for more advanced research towards supervised learning methods to predict sub topics within a text.

3. Managerial Implications

Currently companies are pushing more and more to gather product reviews. These reviews could be a mine of information about all product features. However, when the reviews are not structured it will be hard to get the sentiment about sub topics. This is where this research can help. It allows managers to get more information from the same reviews, which can lead to more customer insight. These in turn could lead to new products which will better suit the needs of the customers. Previous research (Mukherjee & Hoyer, 2001; Archak, Ghose, & Ipeirotis, 2011; Thompson, Hamilton, & Rust, 2005) has shown that product features are very important for the success of a given product. This research can quantify the beliefs about certain product features from reviews. Which can help determine new product features for future products.

4. Limitations and Directions for Future Research

In this research I have used a DTM as independent variable. This makes analysing the data a lot easier. However, this takes out a lot of the text. For example, it would be impossible to detect sarcasm. I also stem the words which could remove superlatives, for example higher and highest will both become high. Getting more info out of the text would require more advanced methods there are previous researches (Hogenboom, 2015) which are able to keep the text structure. This could more accurately predict the sub topic score, which would make the methods more likely to use in actual product development.

Secondly, I have used one data set for the subtopic prediction. When using more datasets, the outcomes could be more robust and applicable to new data sources. This could also be a solution for the inability to predict lower scores in most of my models. All the methods I am using for the prediction are 'out of the box', I did not customize any machine learning programs. In previous papers, which are discussed in the theoretical part, more customized methods where used. I believe these could also lead to better results in the score prediction.

Thirdly, when doing sentiment analysis, you never know if the text really represents the score which it was given. It could be possible that the person who wrote the review, did not write down all his thoughts. This could mean that the score and the review do not align, which makes it harder for the model to predict the value. This will always be the problem with this kind of research, unfortunately I have not found previous research which quantifies this problem.

Fourthly, I split the reviews in paragraphs to determine sub topics per paragraph. In real life there will be very small reviews with only a couple of lines. My method will not be able to determine the sub topics of these small reviews. Making this possible instead of analysing the paragraph the sentences have to be analysed separately. This would require methods like Hogenboom (2015), these are outside the scope of this thesis.

Lastly, I have done no research to the effect of sub topic scores and sales of a product. This method allows to predict the sub topic scores for a lot of reviews which could be used to research the effect between sub topic scores and sales of a product. This could also aid managerial descision towards new product development.

Reference

Annette, M., & Grzegorz, K. (2008). A Comparison of Sentiment Analysis Techniques: Polarizing Movie Blogs. *Advances in artificial intelligence*, 25-35.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science, 57*(8), 1485-1509.

Berger, J., & Milkman, K. L. (2012, April). What Makes Online Content Viral? *Journal of Marketing Research, 2*(49), 192-205.

Bishop, C. M. (1995). *Neural networks for pattern recognition.* Oxdord: Clarendon Press.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (Vol. 4). New York: springer.

Chang, C. C., & Lin, C. J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*(2.3), 27.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research, 43*(3), 345-354.

Chrysanthos, D., Zhang, X. M., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing, 21*(4), 23-45.

Clifton, C., & Robert, C. (1999). TopCat: Data mining for topic identification in a text corpus. *Principles of Data Mining and Knowledge Discovery*, 174-183.

Cui, H., Vibhu, V., & Datar, M. (2006). Comparative experiments on sentiment classification for online product reviews. *AAAI., 6*, 1265-1270.

Dell. (n.d.). *Naive Bayes Classifier - Statistics Textbook*. Retrieved 12 3, 2015, from Dell: http://documents.software.dell.com/Statistics/Textbook/Naive-Bayes-Classifier

Dhillon, I. S. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 269-274.

Dimitriadou, E., Hornik, K., Meyer, D., & Weingessel, A. (2008). Misc functions of the Department of Statistics (e1071). *R package*, 1-5.

Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features., *12*, pp. 194-202.

Entertainment Software Association. (2014). *2014 Essential Facts About the Computer and Video Game Industry.*

Eric , B. T., & Lee, T. Y. (2011, October). Automated Marketing Research Using Online Customer Reviews. *Journal of Marketing Research, 48*(5), 881-894.

Feinerer, I. (2008). A text mining framework in R and its applications. *Doctoral dissertation, WU Vienna University of Economics and Business*.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning. Berlin: Springer.

Gaudiosi, J. (2015, January 15). *Mobile game revenues set to overtake console games in 2015*. Retrieved December 11, 2015, from Fortune: http://fortune.com/2015/01/15/mobile-console-game-revenues-2015/

Ghiassi, M., Skinner, j., & Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications, 10*(16), 6266-6282.

Hogenboom, A. c. (2015). Sentiment Analysis of Text Guided by Semantics and Structure.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Berlin, Heidelberg: Springer.

Karatzoglou, K., Meyer, D., & Hornik, K. (2005). Support vector machines in R.

Kuhn, M. (2015, August 6). A Short Introduction to the caret Package.

Li, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. *Proceedings of the 18th ACM conference on Information and knowledge management*, 375-384.

Ludwig, S., de Ruyter, K., Friedman, M., Brüggen, E. C., Wetzeis, M., & Pfann, G. (2013). More Than Words: The Influence of Affective Content and Linguistic Style Matches in Online Reviews on Conversion Rates. *Journal of Marketing, 77*(1), 87-103.

Moore, D. S., McCabe, G. P., Duckworth, W. M., & Alwan, L. C. (2009). Business Statistics. New York: W.H. Freeman and Company.

Mukherjee, A., & Hoyer, W. D. (2001). The effect of novel attributes on product evaluation. *Journal of Consumer Research, 28*(3), 462-472.

Newzoo. (2015, May 18). *US and China Take Half of $113Bn Games Market in 2018 Read more at http://www.newzoo.com/insights/us-and-china-take-half-of-113bn-games-market-in-2018/#V8W6IgJ0tZm2tbR0.99*. Retrieved December 11, 2015, from Newzoo: http://www.newzoo.com/insights/us-and-china-take-half-of-113bn-games-market-in-2018/

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 1-135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment Classification using Machine Learning. *Association for Computational Linguistics*, 79-86.

Park, C. W., Milberg, S., & Lawson, R. (1991). Evaluation of brand extensions: the role of product feature similarity and brand concept consistency. *Journal of consumer research*, 185-193.

Peng, X., Wu, W., & xu, j. (2011). Will You Be in Hospital Next Year: Leveraging Machine Learning in Improving Healthcare.

Popescu, A.-M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. *Natural language processing and text mining*, 9-28.

Quinlan, R. J. (1986). Induction of decision trees. *Machine learning, 1.1*, 81-106.

R Core Team. (2015). R: A Language and Environment for Statistical Computing. *https://www.R-project.org/*. Retrieved from https://www.R-project.org/

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks , 2*(6), 459-473.

Senecal, S., & Nantes , J. (2004). The influence of online product recommendations on consumers' online choices. *Journal of retailing, 80*(2), 159-169.

Simonson, I., Carmon, Z., & O'curry, S. (1994). Experimental evidence on the negative effect of product features and sales promotions on brand choice. *Marketing Science, 13*(1), 23-40.

Smola, A., & Vishwanathan, S. V. (2008). In *Introduction to machine learning.* UK: cambr idge un ivers ity press.

Taboada, M., Brooke, J., Tofiloski, M., Volt, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational linguistics, 37*(2), 267-307.

Therneau, T. M., & Atkinson, E. J. (1997, Sep 7). *An introduction to recursive partitioning using the RPART routines. Technical Report 61.* Retrieved from nabble.com: http://r.789695.n4.nabble.com/attachment/3209029/0/zed.pdf

Thompson, D. V., Hamilton, R. W., & Rust, R. T. (2005). Feature fatigue: When product capabilities become too much of a good thing. *Journal of marketing research, 42*(4), 431-442.

Tianfeng, c., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature. *Geoscientific Model Development , 7.3*, 1247-1250.

Tirunillai, S., & Tellis, G. J. (2014, August). Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation. *Journal of Marketing Research, 4*(51), 463-479.

Vapnik, V. N., & Vapnik, V. (1998). Vapnik, Vladimir Naumovich, and Vlamimir Vapnik. Statistical learning theory. New York: Wiley.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, 189-196.

Zhang, H. (2004). The Optimality of Naive Bayes. *AA, 1*(2), 3.

Zhu, F., & Zhang, X. M. (2010, March).). Impact of Online Consumer Reviews on Sales: The Moderating Role of Product and Consumer Characteristics. *Journal of Marketing, 2*(74), 133-148.

Zhu, X. (2005). Semi-supervised learning literature survey.

Appendix A Results

Variable		GameReview	Gaming target
Observations		934	2183
Total score			
	Mean	7,54	7,78
	Median	7,80	8,10
	SD	1,43	1,49
	Max	9,80	10,00
	Min	1,00	0,30
Presentation/visuals			
	Mean	7,56	7,86
	Median	8,00	8,00
	SD	1,57	1,48
	Max	10,00	10,00
	Min	1,00	0,00
Gameplay			
	Mean	7,58	7,86
	Median	8,00	8,00
	SD	1,53	1,67
	Max	10,00	10,00
	Min	1,00	0,00
Value			
	Mean	7,31	
	Median	7,50	
	SD	1,75	
	Max	10,00	
	Min	1,00	
Features			
	Mean		7,64
	Median		8,00
	SD		1,68
	Max		10,00
	Min		1,00
Audio			
	Mean		7,71
	Median		8,00
	SD		1,63
	Max		10,00
	Min		0,00

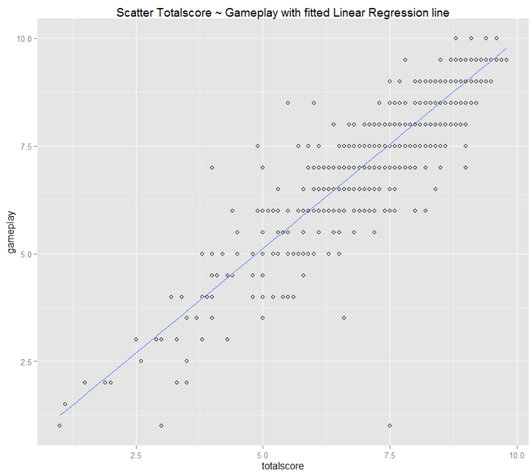
Table 11 Summary of gameReview and Gaming target

	Total score	Visuals	Audio	Gameplay	Replay	Features
Total score	1,00	0,75	0,73	0,91	0,86	0,83
Visuals	0,75	1,00	0,70	0,62	0,55	0,57
Audio	0,73	0,70	1,00	0,59	0,54	0,55
Gameplay	0,91	0,62	0,59	1,00	0,79	0,73
Replay	0,86	0,55	0,54	0,79	1,00	0,80
Features	0,83	0,57	0,55	0,73	0,80	1,00

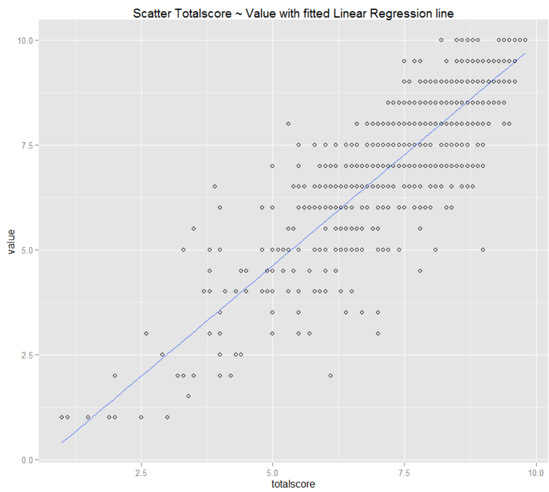
Table 12 Correlation matrix Gaming Target

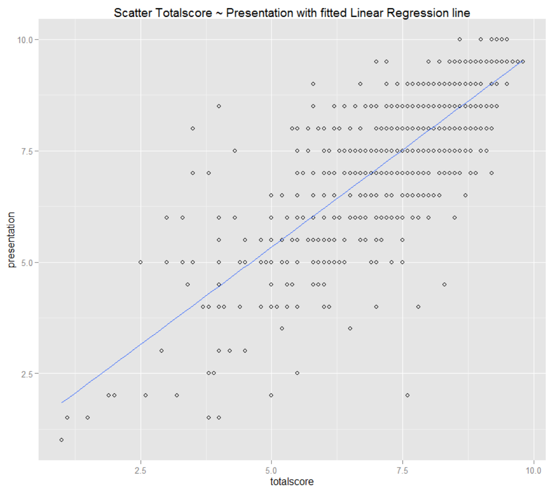
	Total score	Presentation	Gameplay	Value
Total score	1,00	0,79	0,90	0,86
Presentation	0,79	1,00	0,61	0,54
Gameplay	0,90	0,61	1,00	0,73
Value	0,86	0,54	0,73	1,00

Table 13 correlation matrix GameReview



Graph 8 GameReview totalscore ~ gameplay





Picture 1 An example page of Gaming Target

Appendix B technical realisation

For the training and prediction, I use R. All the packages I used can be seen in Table 14. As a SVM package I use the e1071 package from the Technical University of Wien. This package gives R access to the libsvm by Chih-Chung Chang and Chih-Jen Lin. To train the SVM model I use the 'Caret' package. This packages makes it easy to train the SVM model with various values for C and Gamma[2] to see which values gives the best accuracy. The 'Caret' package in combination with the 'doParalell' package gives R the possibility to run code multi core. Because the dataset consists of a huge number of variables, every single word is a variable, meaning that doing computations can take a lot of time. With the 'doParalell' package the computations are processed in a timelier manner. I also use the same package and method to train the Naïve Bayes Models. For training the decision tree I use the Rpart package (Therneau & Atkinson, 1997).

As I did with the subject prediction I will use R to perform the computations for sub topic score prediction. There are various packages in R for predicting neural networks. I have chosen to use the RSNNS package, which gives R access to the 'Stuttgart Neural Network Simulator'. This is a simple to implement and highly customizable neural network library. The package allows to create multiple layer networks, various different training methods and other tuning parameters. Moreover, the package is also included in the 'Caret' training package. As stated in the previous part the 'Caret' package is used for tuning the parameters of the model, to find the best fitting method and to use multiple processor cores. Unfortunately, the standard 'Caret' package can only be used to tune the model with one hidden layer in the network. However, the open source policy of the package has allowed me to alter the package to allow for multiple hidden layer. These adjustments are now included in the latest versions of this package. Lastly, the computations take quite a lot of time to complete on a normal laptop even when using the multi core package. This is very undesirable because I want to test various models for three different models (Audio, Gameplay and Visual) and different subject defining methods. Therefore, I used the cloud service Azure of Microsoft. To perform the calculations, I used a virtual machine with 8 cores and 14 GB of RAM. This machine uses Ubuntu and used Rserver to give a web

[2] I am the author of this feature which is now available in the CRAN package

interface to Rstudio. This allowed me to perform a lot of intensive computations in a relatively short time.

Package	**use**
E1071	SVM and Naive Bayes training
RSNNS	Neural network training
Rpart	Decision tree training
Caret	Parameter tuning for all training methods
DoParallel	Enable R to use multiple processor cores

Table 14 Used R packages