# Two Thumbs Up!

# Sentiment Analysis in Film Reviews

BA Thesis English Language and Culture

Utrecht University

By Denise van Kollenburg

Student No: 36483385

Supervised by

José de Kruif

Simon Cook

June 2014

## Table of Contents:

## 1: Introduction

The field of humanities is undergoing an evolution that may drastically alter the methods of its research. The emergence of big data, digital repositories, and digital tools has presented both new possibilities and new difficulties to the humanities researcher. Big data offer a wealth of information – but only to those who can work with them. The amount of digital data was estimated at 1500 exabytes in 2013 – $10^{18}$ bytes, which, if burned on discs, would form five stacks to the moon – and it doubles every three years (Cukier and Mayer-Schoenberger). The magnitude of digital content calls for a move away from manual sampling. "Where big data is involved […] making a selection through a sampling method becomes increasingly problematic because the end selection always needs to be manageable for an individual researcher" (Eijnatten, Pieters and Verheul 59). In a database of eight million texts, selecting only 1% of the texts would yield 80,000 results. In such a case, both the questions of manageability and representativeness arise.

Luckily, new techniques are emerging that allow even individual researchers to manage these big data sets. One way to derive specific information from digital repositories without reading all the texts is through semantic text analysis. Latent Semantic Analysis (LSA) can extract the meaning of words from large amounts of data, with no human interference – a process called "unsupervised learning", expanded on in the following section (Landauer, McNamara and Dennis). This process has opened up another technique for analysing and interpreting mountains of data: that of sentiment analysis, also called opinion mining. This model for interpretation is used most in evaluating opinion-heavy databases, with texts like customer reviews. Sentiment mining tools can quickly compare and disclose positive and negative attitudes, a promising development for several academic fields as well as for customer-orientated businesses.

The field of sentiment analysis is very dynamic, as is evinced by the multitude of articles – nearly 400 – that were published on the subject in as little as four years. Techniques are rapidly improving and some of the major issues – several of which are discussed in the theoretical framework – are addressed increasingly often. It is worthwhile, therefore, to revisit older articles that described certain impossibilities of machine analyses. In 2002, Bo Pang, Lillian Lee and Shivakumar Vaithyanathan attempted to categorise film reviews by counting the occurrences of several indicator words. After they rejected this method, they applied three common text categorisation algorithms[1] to the problem of sentiment analysis. Due to the speed of developments in the field, some of the problems Pang, Lee, Vaithyanathan and their contemporaries probably encountered – identifying sarcasm or thwarted expectations, for example – may be well on the way to being solved. It will be interesting, therefore, to study these problems and see how the results of sentiment analysis have changed accordingly. After all, the sooner sentiment analysis can be used by researchers to answer their questions, the better. This is why Pang, Lee and Vaithyanathan's dismissal of predefined wordlists in the field of sentiment mining will be tested with some of today's technology. How does recent software compare to the simple find-and-tally method used in 2002? To answer this question, an introduction into the digital humanities and the state of the field of sentiment analysis is first provided. After that, the replication of the original research is described and the results discussed.

[1]: "[A] series of mathematical steps, especially in a computer program [sic], which will give you the answer to a particular kind of problem or question" (algorithm). Simply put, an algorithm is a step-by-step instruction for a machine to perform a certain task.

## 2: Theoretical Framework

In the following section, a concise explanation of the (dis)advantages of using digital tools in humanities research is provided. Next, several key instruments in digital humanities research are introduced and explained, in order to clear the path towards an explanation of sentiment analysis. Some problems are clarified using examples from the International Movie Database reviews – where possible these were taken from the dataset explained in section 4.1.

### 2.1: An introduction to Digital Humanities and its issues

After centuries of humanities research done through close reading of whichever texts were available by chance, it is to be expected that the field will change in the face of a digital revolution. The availability of countless amounts of texts from anywhere in the world will unlock research questions that were impossible to resolve before. For example, new software analysing patterns in writing is achieving great results in identifying who wrote anonymous texts – something that has eluded scholars for ages. The advantages of doing research in the digital age do not need much attention: it has never been easier for researchers to find and access what they are looking for. Big datasets, however, come with several issues that one needs to be aware of. The digitization of analogue texts is not the only actor in the rapid growth of the online collection of texts: the common man plays an even greater part in the expansion. Consider the tweets, Facebook posts, forum contributions and blogs on any subject that are being published – and made researchable – everywhere and all the time, and the need for a change of tactics becomes clear. The field of humanities will have to change the methods they have used for many years to be able to bear these developments.

When analysing a novel, one is generally sure that what they are reading is exactly what they are supposed to read. The text has been proofread by several people and the layout was carefully planned. Researchers may still differ on their interpretation of a certain passage, but more often than not the text they have read is certain to be accurate. Of course, there are exceptions – the phrase "blue-eyed hag" comes to mind[2] – but in these cases the variations are often the topic of research. When analysing big data – texts that are often automatically "scraped" (extracted) from the internet – there is no certainty that the message and the form of the texts is represented correctly. Common issues are the exclusion of search results in which the search term is misspelled, the inability of software to read certain characters, and a misrepresentation of the text's layout. This may lead, for example, to the author's name appearing in the actual text field, and the text being put into a different column and thus excluded from analysis. These problems cause incomplete or tainted datasets, and cleaning them up takes time and money.

For researchers who need to find specific sections in large amounts of data, a cluttered dataset may be a significant impediment. However, most research benefits from quantity over quality: "inaccuracy can be tolerated, because the benefits of using vastly more data of variable quality outweigh the costs of using smaller amounts of very exact data" (Cukier and Mayer-Schoenberger). The discomfort of missing out on some texts is easier to bear if these constitute only a very small part of the entire dataset. A fine example of this is the free software Google Translate offers. The software analyses large quantities of data – "billions of texts" (Cukier and Mayer-Schoenberger) – from the internet, and produces better translation results than translation software that only uses clean data from government documents and dictionaries. Other researchers, such as Landauer and Dumais, also indicate that experience trumps careful assessment of sources (214). Still, digital researchers must always be critical of their results, especially when the texts were analysed automatically. It is vital that researchers are aware of the contaminations in their datasets. Similarly, it is important that researchers gain insight into why certain results are provided. The "black box" of technology – where the

[2] In William Shakespeare's play *The Tempest*, the term blue-eyed hag sometimes turns up as "blew-eyed hag" or even "blear eyed hag", changing the meaning of the passage. Shakespearians do not agree on which term is the original, and it seems it cannot be retraced. (Waters)

user does not understand the process that generates the output – is a problem that needs to be diminished as much as possible.

The shift from analogue to digital research has another effect which a researcher must be aware of: that of representativeness. Before computers were involved, researchers were generally aware of what they were – and more importantly, were not –studying. In big databases, it is difficult to discover what is not included. This can be illustrated by the development of online archives: a crucial obstruction in digitizing texts is copyright, and it is especially troublesome regarding twentieth century content. The copyright issues skew the representativeness of – for example – newspaper collections; local newspapers will be more eager to share their publications than large, profit-based newspapers. Besides, if the *Daily Mail* were to donate their entire archive, and *The Guardian* would not, this would yield different results than if the opposite situation had occurred. That which is not digitally available is therefore just as significant as what *is* available online, but it is far more difficult to pinpoint. The problem with this "offline penumbra" is stressed by Patrick Leary: "The time is near upon us when whatever is not online will simply cease to exist as far as anyone but specialists is concerned" (82). It cannot be counted upon that everything a researcher needs is available online.

This is only a selection of the issues that arise with the new techniques. If the scientific world were to fail to address them, many uses of digital tools and texts would go undetected. As of yet, those humanities scholars who have embraced the digital world are still outnumbered by those who do not fully accept all the digital age has to offer. However, Prescott warns the devoted supporters of the digital humanities: "One of the problem[s] confronting data enthusiasts in the humanities is that we feel a need to convince our more old-fashioned colleagues about what can be done. But our role as advocates [of] data shouldn't mean that we lose our critical sense as scholars" (Prescott). Ideally, awareness of both the advantages and the disadvantages will be spread, so that big data research can grow to become a fundamental part of the humanities.

## 2.2: Towards Sentiment Analysis

There are many fields in which sentiment analysis software can be of great help: it is in many cases impossible to read and manually assess the nature of all relevant texts. Fields like history, media studies, gender studies (and many more) could benefit greatly from using sentiment mining tools to automatically generate statistics to support or disprove their claim. Similarly, businesses can quickly gain awareness of which problems arise with their product, and which aspects are especially praised. This makes research into improving sentiment analysis a lucrative industry, since it is vital that the software returns as many results as possible. After all, any reviews that are *not* found and addressed by the business might still be found by users who are looking to buy a product – and who could refrain from such a purchase because of negative reviews. Failure to address complaints could discourage customers. Before sentiment analysis could developed, however, the following aspects needed to be addressed.

### 2.2.1: Supervised and unsupervised learning

Sentiment analysis is a form of classification. Taking a database of film reviews as an example, these reviews can be sorted either into positive or negative attitudes, or sorted on the topic discussed – such as acting, plot, or special effects. This can be done through machine learning. A machine learning algorithm differs from a regular algorithm in that it will learn which steps to take through examples. There are two ways through which software can be taught to produce the correct output from a certain set of data, both explained by Zoubin Gharmani. In supervised learning, the software is not only given the input data, but also a sequence of desired outputs, to teach the machine to produce the correct output. In the case of the reviews, for example, the machine would be given a set of reviews and would be told which reviews were positive and which were negative. The machine would apply the knowledge it gained from this process to unlabelled datasets: for example, it may deduce that the word “excellent” was mentioned more often in positive reviews. The measure in which software is trained can vary, from only providing it with minimal instruction to teaching it in detail what output must be produced. In unsupervised learning the machine is not given any indication of what output it should generate. This may seem like a gamble: after all, a machine cannot understand what it is processing, so how can it decide what to do with the data? It should be realized that the machine will attempt to represent the data so it can be used for “decision making, predicting future inputs, […] communicating the inputs to another machine, etc.” (Ghahramani 3). The machine will look for patterns in the data, and one way of structuring these data – a useful one in sentiment analysis – is through clustering: it will assemble a specified or unspecified number of groups in which the data fit. It might form a cluster of reviews containing the words “excellent” and “acting”, and a cluster of the words “awful” and “plot”. These clusters would then be ready to be classified by positive and negative sentiments, if such a sorting were required.

### 2.2.2: Natural Language processing

The difference between natural language and computer language could not be greater. While computer language is straight-forward and explicit, natural, human language is full of ambiguities and inconsistencies. Before automatic analysis can take place, this language must be translated into text representations which can be used by computing systems. “At the core of any NLP task is the important issue of natural language understanding” (Chowdhury 55). Several techniques already exist to prepare regular texts for automatic analysis. These techniques vary between finding meaning on word level, sentence level, and even the meaning in and of the entire context. First, the texts may be separated into High Frequency Words and Content Words (Tsur, Davidov and Rappoport 164). When a computer needs to understand the message of a text, an article does not carry the same weight as a noun. Depending on the type of text analysis, words that occur less frequently could be more significant in the text analysis. This statistical approach is called Term Frequency – Inversed Document Frequency (TF-IDF) by Rajaraman and Ullman (8). Another technique on word level is the stemming of words to reduce the morphological variety. When the term "actor" is searched, the machine will also return results like "acting", "act" and "actress", because all these words can be returned to the stem "act". This will simplify and speed up the search process significantly. Part-of-speech tagging can also be carried out to further clarify the meaning of a word, and so avoid confusion over whether "act" is a verb or a noun. Extending on this is the automatic parsing of the entire sentence, and thus separating different clauses. This can be especially useful in sentiment analysis, as one can understand after reading the following sentence: "He not only looked the part of the Green Goblin, but brought some welcome gravitas to the role" (TheLittleSongbird). This one sentence mentions two different aspects of the actor – the first being "looks" and the second "the ability to act" – and these might in some cases be marked separately. Another task in Natural Language Processing that is applicable to the field of sentiment mining is that of Named Entity Recognition. If the software understands it is dealing with a named entity – for example, an actor's name or a place name – rather than seeing it as a random noun, its chances of labelling the entity correctly might improve. These and several other processes can be used simultaneously to allow machines to interpret texts successfully.

### 2.2.3: Semantic Text Analysis

Semantic Text Analysis engages in the actual deriving of the meaning of words and sentences. The meaning of a word is derived from the words that are in its semantic space – words get their meaning indirectly, from the relation they hold with other words (Landauer, McNamara and Dennis 91). The example the researchers provide is this: the chance that the word “semantic” and “carburettor” are used simultaneously is slim, as the words are completely unrelated. Similarly, the words “gears” and “brakes” might not be used together because they identify two aspects of the same object. Through semantic text analysis, software might discover that “gears” and “brakes” both occur in texts with the word “car”, but “semantic” and “carburettor” have no such link. LSA analyses the texts and thus can come to an automatic understanding of the words through *experience* – a central concept in LSA. The learning power of the LSA model was tested and compared to the rate at which school-aged children improved their performance – i.e. the extent of their vocabulary – through reading. The model approximated the natural rate of improvement of schoolchildren, an interesting development since schoolchildren seem to learn by many other ways than just adult instruction (Landauer and Dumais 2). Teaching a machine to understand what a text is about is, of course, an indispensable step towards automatic sentiment analysis.

## 3: Sentiment analysis

Sentiment analysis can deal with simply positive and negative categories (while sometimes a neutral class is added, it brings its own problems – as is discovered in section 5), as well as provide a ranking system through which the polarity of opinions is determined. Unfortunately, sentiment mining comes with many inherent difficulties. Most texts do not come with a built-in ranking system for sentiments. Teaching a machine the meaning of words is one thing, making it understand the mood of a sentence or text is another. Ronen Feldman identifies several different areas for sentiment analysis. For example, analysis can be carried out at document level, at sentence level (including phrases), and analysis can be aspect-based or feature-based. In this situation the software recognises opinions expressed in a document and links them to the relevant aspect (83). Since it is likely that the field of sentiment analysis has seen promising developments in twelve years' time, recent progress in some of the major fields is elaborated on. Much may be changed since Pang, Lee and Vaithyanathan conducted their experiment in 2002. To ensure relevance in the discussion of the state of opinion mining, only papers on sentiment mining problems that date from 2010 or later are included.

## 3.1 Thwarted Expectations

It is unmistakably true that any type of ambiguity provides a challenge for automatic sentiment analysis. For example, one could imagine that a machine would have problems identifying the following review as positive:

> Having seen the trailers for this film I have to say that I didn't walk into the cinema with high hopes. The computer effects looked badly integrated, the Green Goblin's costume looked awful and comic book adaptations usually have such painful scripting and plotting. Thankfully I was wrong on most counts. (to_kill_better)

No human reader would have trouble understanding that the reviewer enjoyed the film. What marks the review as positive, however, is the word "wrong". The above review belongs to what Pang, Lee and Vaithyanathan identify as a "thwarted expectations" review. This is defined as a review "where the author sets up a deliberate contrast to earlier discussion (85). If software only takes separate features into account, it would find too many negative words to classify it as a positive review. Pang, Lee and Vaithyanathan suggest "some way of determining the focus of each sentence" is found, "so that one can decide when the author is talking about the film itself" (85).

In the same vein of misunderstanding lies the ambiguity created by sentiment shifters and negation words. The solution might seem straightforward: any opinion word accompanied by a negation simply switches from one end to another. However, consider the following sentence: “epic movie was kinda fun but not as good as disaster movie” (Danny DW4). A human reader understands that the claim that is negated is “as good as disaster movie”. For a machine, however, it is just as likely that *Epic Movie* is not fun at all. Words like “never” and “nobody” might be even more difficult to evaluate correctly, as would the sentiment shifter “but” in the following judgement: “Orlando Bloom […] I think he’s good looking but that’s it really” (teresa6). The opinion that Bloom brings nothing to the screen except for attractiveness is – for the moment – hidden from most sentiment analysis software. Similarly, “There were also a few corny lines but in a three hr 20 min film this is hardly a major flaw!” (idem) might be understood as a negative sentiment if the word “hardly” is not processed correctly. This problem might also be solved by separating different clauses in a sentence and identifying which clause a negation or shifter refers to.

## 3.2: Implicit Aspects

Feature-based sentiment analysis can be achieved by singling out the noun phrases, and can be used to find out exactly what makes users rate poorly, whether it is the acting or the plot. However, in general, feature-based analysis excludes implicit aspects such as “the film is too long!”, of which the explicit aspect is “length of time”. In some databases, not accounting for implicit aspects can significantly reduce the amount of retrieved relevant texts. In a database of 500 film reviews, 33 of a total of 50 mentions of the word “long” are an implicit aspect of the explicit feature “time”. Only 12 are explicit, the others are part of collocations like “as long as”. In 2011, this problem was addressed by Hai, Chang and Kim, who have developed an approach for solving the problem of implicit aspects.

Co-occurrence association rule mining (coAR) investigates the bonds between implicit features and their explicit equivalents. Through semi-supervised learning (where only a small dataset is used for teaching purposes), their programme creates rules from "frequent associations between all opinion words and explicit features in the corpus. It then applies the rules to each opinion word […] identifying its most likely associated feature" (Hai, Chang and Kim 396). If the software finds several mentions of "it took a long time", it will understand that "the scene took too long" is commenting on the length of time, not on the length of someone's coat. By implicating rules of minimal support, the machine discards pairs of opinion words and explicit features that are only put together in rare cases. "Long" does not usually pair up with "raincoat", but in some reviews of *The Matrix Reloaded* it does. These few instances should not affect the general meaning of the implicit feature and should therefore be left out of consideration.

## 3.3: Sarcasm

Sarcasm is notorious in the field of sentiment mining. Ambiguous by nature – proven by the fact that even humans sometimes fail to identify it – it is especially harmful to opinion mining systems because its explicit meaning is often the exact opposite of the intended meaning. To understand the intended meaning, a degree of world knowledge is often required. In a paper published in 2010, Tsur, Davidov and Rappoport develop an algorithm for identifying and classifying sarcastic statements from product reviews. They first admit that research has already been carried out to automatically identify only single subcategories of sarcasm, but their research focuses on recognising any type of sarcasm from both sarcastic and regular reviews. Although the algorithm is not presented in the paper, the discussion of data acquisition and results reveals the precision rate[3] of 76% and the recall rate[4] of 81% to be

[3] The amount of correct texts retrieved compared to the amount of incorrect texts retrieved.

achieved through pattern recognition. The machine extracted the patterns of a set of labelled sentences – rated from 1 (no sarcasm) to 5 (very sarcastic) – and selected criteria to trim the results. For example, they discarded patterns that occurred both in class 1 and in class 5, as they would certainly not be useful in sarcasm detection. New data were tested on the basis of these patterns. The software that was built (SASI) achieved a precision rate of 77% and a recall rate of 83.1%. Tsur, Davidov and Rappoport recognise certain aspects which often signal sarcasm, such as the use of ellipses, capital letters, and the word "great" when used as a word sentence. However, as these features can also signal other speech acts – surprise or anger, for example – they do not carry much weight (168). Although the success rates mentioned in the article are certainly promising, sarcasm detection remains one of the most elusive parts of opinion mining. More research will have to be performed in order to truly grasp the concept.

## 3.4 The research by Pang, Lee and Vaithyanathan

In 2002, Pang, Lee and Vaithyanathan published a paper on classifying documents by overall sentiment: "Thumbs up? Sentiment Classification using Machine Learning Techniques". They state that before their research, no unsupervised machine learning research had been performed in the field of sentiment mining. In earlier research – by Hatzivassiloglou and Wiebe, for example –the techniques had only been used for detecting opinion texts, not for classifying them. Pang, Lee and Vaithyanathan chose to use user-generated film reviews for their research, as these often contain extractable rating indicators in the form of stars or a numerical value (80). In a preliminary test, they discredited the idea that sentiment analysis can be done by lists of words frequently used in texts containing sentiments. They stated that, to perform sentiment analysis, it is not enough to simply search texts for indicator words.

[4] The amount of correct texts retrieved compared to the amount of correct texts in the entire database.

Instead, they vouched for machine learning: “the machine learning algorithms clearly surpass the random-choice baseline of 50%[5]. They also handily beat our two human-selected unigram[6] baselines of 58% and 64%”. The unigrams mentioned in this quotation are words like “brilliant”, “gripping” and “unwatchable”, selected by students to designate positive or negative film reviews. It was, however, admitted that the list of words they used may not have been optimal (80-81). Pang, Lee and Vaithyanathan simply counted the occurrences of these words in the reviews and used this to predict the class in which they would likely be placed. The number of reviews that were rated equally likely was as high as 75% for one word list. The researchers compiled a list based on frequency counts in the entire corpus and compiled a new list of fourteen terms. This was done to investigate the role of the length of the word list– or lack of it. Accuracy rose only marginally (to 69%) but the tie rate dropped to 16%[7]. The substandard results led them to applying machine learning techniques: feeding the software texts containing opinions and leaving the machine to categorize these opinions automatically yielded far better results. Three different algorithms commonly used in text classification were applied to a database of 1,400 film reviews. Success rates ranged from 72.8% to 82.9%.

It will be interesting to investigate how far the research would have come in the field that Pang, Lee and Vaithyanathan readily dismiss: that of using a pre-defined word list. The researchers have themselves recognised that the word list assembled after careful examination had a positive effect on the accuracy. There are many pre-constructed word lists available for sentiment analysis, and although they are not created specifically for film reviews, they might still yield decent results.

[5] A baseline is “a value or starting point on a scale with which other values can be compared” (Baseline). If a machine were to randomly assign reviews into one of two classes, the chance it would classify one correctly would be 50%. Anything over that surpasses the random-choice baseline.

[6] Unigrams, as opposed to bigrams or ngrams, are simply single words.

[7] The table with the word lists and their accuracy rates and tie rates can be found in the appendix (fig 1).

## 4: Methodology

In order to test and demonstrate the usability and accuracy of some current sentiment mining techniques, a case study as performed by Pang, Lee and Vaithyanathan was reconstructed. The goal of this case study was to evaluate if the results achieved in the research from 2002 could be improved upon, or whether using word lists to analyse sentiments was still an unrewarding process.

## 4.1: Compiling the Dataset

A dataset of 500 film reviews was constructed from a randomly selected number of films (the list of which, including some relevant metadata, can be found in the appendix (fig 2)). These were mined from the user reviews section of the International Movie Database (IMDb) using the Google Chrome Web Scraper tool. The International Movie Database was the same source as was used by Pang, Lee and Vaithyanathan, although most likely a completely different set of reviews was extracted. Because of this, if any significant changes in language had occurred since 2002, these might influence the results. Due to time limitations, however, this problem was not explored any further.

Only reviews with in-text ratings of the type *x*/10 were used, other reviews were discarded. Due to the nature of the reviews (user-generated), all ratings were made comparable. They were rounded up to single number numerators manually (to avoid confusion about ratings like 3.5/10, which might be read as 5/10) and a denominator of 10 (in order to nominalise ratings of *x*/100). Any ratings of zero and all negative ratings (six negative ratings in total, ranging from "-infinity/10" to a tolerant -2/10) were saved as zero. For the sake of manageability and comparability, the list was trimmed to 200 positive reviews (rated 8, 9 or 10), 100 neutral reviews (6 or 7) and 200 negative reviews (rated below 6). These ratings were extracted to another column for evaluation purposes. Another column indicated which film was reviewed. Lastly, the reviews were given a unique identification number so that the original class they were rated in could be identified: 1 through 200 were rated positive, 201 through 300 were rated neutral, and 301 through 500 were rated negative. The file was saved as a Comma-Separated Values document. To ensure the columns were read properly, all commas, semicolons, quotation marks (single and double), enters and double spaces were removed from the review text.

## 4.2: The Software

For the research, SPSS Modeler [sic] 14.2 was used, which was published in 2011 by the International Business Machine Corporation. The IBM® SPSS© Text analytics add-on tool was used to carry out the analysis of the database.

First, the csv file was uploaded and the settings were altered according to the format used. After adding a Text Mining node to the uploaded file, the template for English opinions was used to extract concepts of several types. These included but were not limited to positive and negative feelings, budget, attitude, and competence. After assuring the settings yielded the desired results, a model of these concepts was generated automatically. Two separate tables were created as csv output, respectively scoring the reviews on occurrences of positive and negative concepts. These tables were further analysed using Microsoft Office Excel.

## 4.3 Collecting the results

Following is a report of the further processing of the positive concepts table. Similar actions were performed on the negative table. First, the number of mentions of all concepts were added up and displayed per review. Secondly, the total number of words per review was calculated. By dividing the number of positive sentiments by the total number of words, an indication of the overall proportion of positive sentiment per review was calculated. After sorting this indication from largest (= relatively many positive terms) to smallest (=relatively little positive terms), the wrongly classed reviews were identified. Ideally, the 200 reviews with the (relatively) highest amount of positive words were also the reviews given an ID between 1 and 200, etc. A review was classified incorrectly if the ID number assigned before analysis no longer matched the place in the list after analysis. All the incorrectly classified reviews were marked, and for each rating (0-10) it was counted how many out of the total were correct.

In another table, the number of reviews that were correctly classified was calculated in percentages for each rating. Based on these results, it was decided to incorporate the neutral class into the positive class. The reason for this will be related in section 5.2. This was done by allowing the first 300 results to be marked positive. These results were also added to the table. The average number (both positive and negative combined) of correctly classed reviews was added to the table as “Total Score”.

# 5: The Research

## 5.1: On the usefulness of the SPSS software

It should be noted that SPSS Text Analytics is not updated with state-of-the-art technology, as most of the algorithms and programmes discussed earlier have not (yet) been made public. The newest algorithms are not generally available due to copyright restrictions. Instead, the tool should be seen as a representation of the state of the field. SPSS seems to choose offering more automatic text analysis methods rather than offering a select few of the very best. It can therefore not be expected to solve the problems of sarcasm and thwarted expectations. Since SPSS Text Analytics is available for purchase, however, it is very likely that the software will be used by many researchers who do not have the funds to purchase or knowledge to build superior software. Because of this, the following research can be relevant to identify problems that may commonly occur in sentiment research.

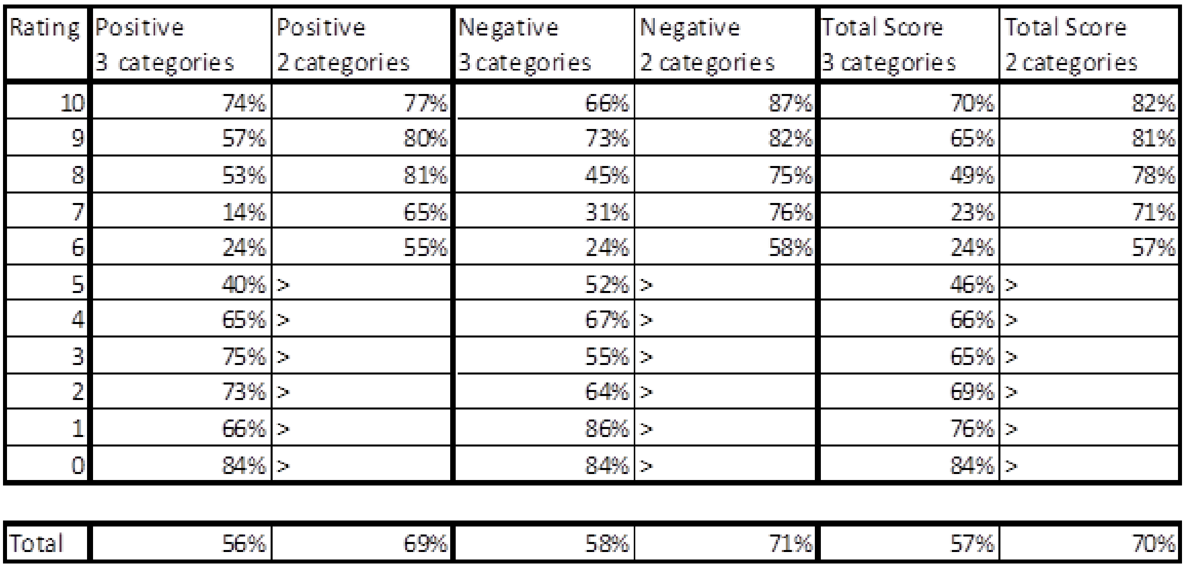
The SPSS Opinions (English) template is little more than a predefined word list, which can be applied to whichever text that needs to be analysed automatically. The word list is, however, far advanced – especially in quantity – from the word list used by Pang, Lee and Vaithyanathan: "we asked two graduate students in computer science to (independently) choose good indicator words for positive and negative sentiments in movie reviews" (80). These two word lists achieved 58% and 64% accuracy. After the researchers inspected some of the source material, their updated word list reached an accuracy of 69%. Since such an improvement took place after only a minimal inspection of the data, the larger and more developed English opinions word list in SPSS is worthy of examination. Due to its size, this list cannot be examined in full. Only indications can be found of how the list produces the results. It must therefore function as a black box, where only the input and the output are known. Whether the software is more adequate than the 2002 word list will become apparent in the results.

The SPSS word lists include common spelling and typing mistakes and take collocations into account. In a dataset of user-generated content, this is no unnecessary feature: the amount of incorrectly spelled words would prove a substantial problem. Also, rather than marking the phrases "I love this film" and "I really like this film" separately, it makes use of "underlying terms": words that the software deems similar are grouped together under a single term.[8] Beyond that, it makes use of stemming to further minimize the amount of output. This trimming of results makes it easier for researchers to process the data, although it also comes with certain drawbacks. Since the underlying terms are not specifically grouped for analysis of film reviews, there is what is called "fuzzy grouping": words that

[8] An example of a list of underlying terms can be found in the appendix.

should not be similar are still marked together. An example of this is the grouping of “original” and “creative”. It is easy to produce sentences in which these words are far from interchangeable, and indeed, most reviews marked “creative” by the software only mention the original movie when discussing a sequel. Similarly, the software does not recognise “special effects” as a single concept that can be rated positively or negatively, but rather marks it as two concepts: a positive opinion and an unknown concept. The fact that this word list was not created specifically for the domain of films has a negative overall effect on the success of the analysis.

Several concepts are combined with negation words and counted separately. For example, the software recognises “not realistic” and “not interesting” as negative concepts. Unfortunately, not every use of a negation word is marked and thus the problem of negation is not eliminated entirely. Taking in only a few uses of negation merely scratches the surface of the problem and can barely be considered a solution. SPSS also does not take sentiment shifters into account. “Too fast” was grouped under positive concepts, because the software failed to attribute the negative meaning of “too”. Words like “but” were not added to any of the generated concepts, and would not function as sentiment shifters. Some implicit aspects were added to the word list and assigned to a class. “Long”, the word used as an example earlier, was added to the “negative” group. Some other hidden aspects were added to the “contextual” class, which refrained from passing judgement: “small” and “new” are words which are classified as being opinion words but are not rated as either positive or negative. Through the inclusion of some negative phrases and implicit aspects, SPSS seems to address – but not solve – some of the issues mentioned in the theoretical framework. Large amounts of negated or hidden opinions that are not included in this list cannot be sorted correctly, and therefore prove an impediment in the actual sentiment mining process.



| Rating | Positive 3 categories | Positive 2 categories | Negative 3 categories | Negative 2 categories | Total Score 3 categories | Total Score 2 categories |
|---|---|---|---|---|---|---|
| 10 | 74% | 77% | 66% | 87% | 70% | 82% |
| 9 | 57% | 80% | 73% | 82% | 65% | 81% |
| 8 | 53% | 81% | 45% | 75% | 49% | 78% |
| 7 | 14% | 65% | 31% | 76% | 23% | 71% |
| 6 | 24% | 55% | 24% | 58% | 24% | 57% |
| 5 | 40% | > | 52% | > | 46% | > |
| 4 | 65% | > | 67% | > | 66% | > |
| 3 | 75% | > | 55% | > | 65% | > |
| 2 | 73% | > | 64% | > | 69% | > |
| 1 | 66% | > | 86% | > | 76% | > |
| 0 | 84% | > | 84% | > | 84% | > |
| Total | 56% | 69% | 58% | 71% | 57% | 70% |

Table 1: Accuracy results for SPSS word list Opinions (English)

## 5.2 Results and Discussion

The baseline of 69% that was achieved by Pang, Lee and Vaithyanathan was initially not even touched by the SPSS software. Table 1 shows that for the positively rated reviews (rated 8-10) scored by positive concepts, the initial percentages were very diverse, ranging from as little as 55% accuracy – a score only marginally better than the random-choice baseline of 50 % - up to 74% correct. The overall rating is shockingly low: only 56% of the reviews was put in the right class. This percentage is as low as it is because of the poor classification of the neutral reviews: reviews rated 7 were only placed correctly in 14% of the cases, reviews rated 6 in 24% of the cases. In the table scored with negative concepts, the same general pattern can be identified: the neutral class scores much below the others, though not as dramatically as in the positive table. Still, it is intriguing that these ratings score so far below the random choice baseline: in some cases, the software is far more likely to classify the reviews incorrectly. At the moment, it is unclear why this may be the case. The overall score for correctly classified negative reviews is 58%. This brings the total score for both tables to 57% labelled correctly. Since these results are hindered so clearly by the neutral class, a logical next step is that of removing this class.

The neutral reviews – rated 6 or 7 out of 10 – were added to the positive class.

Combining the two classes had an undoubtedly positive effect on the overall accuracy of the automatic classification. Although the reviews in the upper middle of the rating spectrum still achieve less accuracy than the poles, the improvement is evident. This total average score improved from 57% accuracy to 70% accuracy. However, if the neutral reviews were cut out entirely, the accuracy rate would improve much more radically. Although this is what Pang, Lee and Vaithyanathan reported on doing, their article does not specify which ratings were cut from the dataset: "[r]atings were automatically extracted and converted into one of three categories; positive, negative, or neutral. For the work described in this paper, we concentrated only on discriminating between positive and negative sentiment" (80). Some indication will be given as to what the accuracy rates might have been if the neutral class had been cut out entirely. If the neutral class as defined in this research (reviews rated 6 and 7) had been cut out, the total accuracy rate would rise only to 71%. However, if 4 and 5 were counted as neutral as well – a reasonable assumption as this would eliminate the four middle ratings – the accuracy rates would climb up to 76%. Although this would not be helpful in practical applications where neutral reviews do exist, it poses a better comparison to Pang, Lee and Vaithyanathan's results.

It can clearly be observed that the accuracy rises as either pole is approached. This conclusion is not discussed in Pang, Lee and Vaithyanathan's research, but can be drawn from the results table (table 1). It seems easier to automatically classify reviews from users that are either extremely satisfied or dissatisfied. This could be be explained by the assumption that an excited reviewer – either positively or negatively – would be more likely to use a large quantity of expressive words. This could also explain why approaching the neutral classes generally means a decline in accuracy. The software is slightly more likely to classify reviews correctly on the basis of negative concepts. Especially in the polar ends, these accuracy rates are generally slightly superior to the accuracy rates of the positive concepts.

The observed higher accuracy rates of negative reviews may be attributed to the greater-than-thou attitude that Tsur, Davidov and Rappoport mention in their article: “we speculate that one of the strong motivations for the use of sarcasm […] is the attempt to “save” or “enlighten” the crowds and compensate for undeserved hype (undeserved according to the reviewer)” (168). Perhaps this “undeserved hype” could cause extra severe reviews. This hypothesis may be supported by the fact that a large part (27%) of the negatively classed reviews in fact reviewed the film “The Matrix Reloaded”, which had indeed suffered quite a build-up. Several other severely rated films – Spider-Man and Pacific Rim – may be added to this class as well. However, this is no more than a hypothesis: the results of this research are not sufficient to prove the claim. Other reasons may exist as to why negative reviews are easier to classify for the SPSS sentiment analysis software.

## 6: Conclusion

The field of digital humanities, although promising many new possibilities, is not the paradise one might expect it to be. Researchers must be prepared to deal with more uncertainty and messiness – in the form of insufficiently specified and cluttered (meta)data – than they might be accustomed to. Despite that, many new developments suggest great results in analysing texts as well as a diminishing of the problems of digital humanities are being reported on. Machines are increasingly capable of managing, understanding, labelling, and classifying texts, and the rate and quality with which this is happening is remarkable.

Nevertheless, sentiment analysis is still in its infancy. Issues are addressed and sometimes resolved, but these solutions have yet to reach the wider field. The purchasable sentiment analysis tool that was used in this research is average at best. For the regular researcher, classifying sarcastic reviews, reviews voicing thwarted expectations, and reviews containing implicit aspects is still out of reach.

Revisiting sentiment analysis research published in 2002 has proven no revolution has taken place in the field. Pang, Lee and Vaithyanathan achieved an accuracy rate of 69% in their automatic analysis of fourteen concepts. The larger and more thorough word list offered by the SPSS software only produces marginally better results: 70% of the reviews was classified correctly. It is, however, unclear how Pang, Lee and Vaithyanathan eliminated the neutral class, making a comparison unstable. Perhaps the accuracy rating of 76% is more applicable. However, this makes no difference in practical applications, as neutral reviews do exist.

Some of the problems in the field of sentiment analysis may be solved very soon. A great step forward would be the automatic determination of the topic of a sentence or clause. This would simplify the solution to the problem of negation words and sentiment shifters. Beyond that, the field would benefit from openness of codes and algorithms. The problem of copyright restrictions hinders the development of better software, and many research questions are still out of reach because current available sentiment mining tools are not fitted with the latest technologies. Nonetheless, the speed with which sentiment analysis develops is promising, and accuracy rates will likely improve significantly over the next few decades.

## Works Cited

"algorithm." *Collins Cobuild Advanced Dictionary*. HarperCollin Publishers, 2009.

"Baseline." *Collins Cobuild Advanced Dictionary*. HarperCollins Publishers, 2009.

Chowdhury, Gobinda G. "Natural Language Processing." *Annual Review of Information Science and Technology* 31 January 2005: 51 - 89.

Cukier, Kenneth Neil and Victor Mayer-Schoenberger. "The Rise of Big Data." *Foreign Affairs* (2013). Web. 30 April 2014.

DW4, Danny. "good movie!! ignore all the bad saying!!" *Reviews for Disaster Movie (2008)*. International Movie Database, 4 October 2008. web.

Eijnatten, Joris van, Toine Pieters and Jaap Verheul. "Big Data for Global History: The Transformative Promise of Digital Humanities." *BMGN - Low Countries Historical Review* 128.4 (2013): 55 - 77. print.

Feldman, Ronen. "Techniques and Application for Sentiment Mining." *Communications of the ACM* 56.4 (2013): 82-89. Print. 12 May 2014.

Ghahramani, Zoubin. "Unsupervised Learning*." *Gatsby Computational Neuroscience Unit* 16 September 2004. Web.

Hai, Zhen, Kuiyu Chang and Jung-jae Kim. "Implicit Feature Identification via Co-occurence Association Rule Mining." *Computational Linguistics and Intelligent Text Processing* 6608 (2011): 393-404. Document.

Hatzivassiloglou, Vasileios and Janyce Wiebe. "Effects of adjective orientation and gradability on sentence subjectivity." *Proceedings of COLING*. 2000.

Landauer, Thomas K and Susan T. Dumais. "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge." *Psychological Review* 104.2 (1997): 211-240. Print.

Landauer, Thomas K., et al. "Preface." Landauer, Thomas K., et al. *Handbook of Latent Semantic Analysis*. New York: Routledge, 2013. Print.

Leary, Patrick. "Googling the Victorians." *Journal of Victorian Culture* 10.1 (2005): 72-86. Print. 1 May 2014.

Marcus, Leah S. *Unediting the Renaissance*. New York: Routledge, 1996. Print.

Nicholson, Bob. "The Digital Turn." *Media History* 19.1 (2013): 59-73. Print.

Pang, Bo, Lillian Lee and Shivakumar Vaithyanathan. "Thumbs up? Sentiment Classification using Machine Learning Techniques." *Proceedeings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* July 2002: 79-86. 5 May 2014.

Prescott, Andrew. *The Deceptions of Data*. 13 January 2013. Blog post. 15 June 2014.

Rajaraman, A and J.D. Ullman. “Data Mining.” *Mining of Massive Datasets*. 2001. 8.

teresa6. “Wow!!” *User Reviews for "The Lord of the Rings: The Return of the King (2003)*. International Movie Database, 30 December 2003. web.

TheLittleSongbird. “Definitely exceeded my expectations!” *User Reviews for Spider-Man (2002)*. International Movie Database, 8 may 2009. Web. 2014 June 14.

to_kill_better. “Amazingly Spectacularly Great.” *User reviews for Spider-Man (2002).* International Movie Database, 26 June 2002. Web. 8 May 2014.

Tsur, Oren, Dmitri Davidov and Ari Rappoport. “ICWSM - A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews.” *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. 2010. 162 - 169. Document.

Waters, Claire R. “The Tempest's sycorax as 'Blew eye'd hag': A note toward a reassessment.” *Notes and Queries* 56.4 (2009): 604-605.

## Appendix 1: The accuracy tables by Pang, Lee and Vaithyanathan

| | Proposed word lists | Accuracy | Ties |
|---|---|---|---|
| Human 1 | Positive: dazzling, *brilliant, phenomenal, excellent, fantastic*<br>Negative: *suck, terrible, awful, unwatchable, hideous* | 58% | 75% |
| Human 2 | positive: *gripping, mesmerizing, riveting, spectacular, cool, awesome, thrilling, badass, excellent, moving, exciting*<br>negative: *bad, cliched, sucks, boring, stupid, slow* | 64% | 39% |

Figure 1: Baseline results for human word lists. Data: 700 positive and 700 negative reviews.

| | Proposed word lists | Accuracy | Ties |
|---|---|---|---|
| Human 3 + stats | positive: *love, wonderful, best, great, superb, still, beautiful*<br>negative: *bad, worst, stupid, waste, boring, ?, !* | 69% | 16% |

Figure 2: Results for baseline using introspection and simple statistics of the data (including *test* data).

## Appendix 2: Table of films used to compile the dataset of reviews

| ID | Film Title (as on IMDB) | Release | Avg Rating |
|---|---|---|---|
| A | After Earth | 2013 | 5 |
| B | Disaster Movie | 2008 | 1,9 |
| C | Dorian Gray | 2009 | 6,3 |
| D | Eat Pray Love | 2010 | 5,6 |
| E | Grown Ups | 2010 | 5,9 |
| F | The Matrix Reloaded | 2003 | 7,2 |
| G | Ocean's Eleven | 2001 | 7,8 |
| H | Pacific Rim | 2013 | 7,1 |
| I | The Lord of the Rings: The Return of the King | 2003 | 8,9 |
| J | Romeo + Juliet | 1996 | 6,9 |
| K | Spider-Man | 2002 | 7,3 |
| L | Taken | 2008 | 7,9 |
| M | The Wolf of Wall Street | 2013 | 8,3 |

## Appendix 3: The underlying terms of the concept “like” (157 concepts)

Underlying terms of “like”

adept, am open to, apprciate, appreciate, appreciates, appreiate, appriciate, appriciates, big fan of, definite must, definite musts, did like, do like, enjay, enjopyable, enjoy, enjoy very much, enjoyable, enjoyed, enjoyed my time, enjoyed ourselves, enjoying, enjoypable, enjoys, enjpoy, enyjoy, especially like, especially liked, fan of, fancy, favorite, favorites, favourite, fond of, going easy on, greatly like, have always liked, he like, hobbies are, huge fan of, i already like, i also like, i appreciated, i did enjoy, i fancy, i go for, i just like, i just love, i like, i like it better than, i like that, i like that it is, i like the concept of, i like the fact, i like the first one the best, i like the way, i like this one, i like this one best, i like this one the best, i liked, i lke the way, i love, i love it, i love the fact, i love the fact that, i loved, i very much like, i'd love, i'm open to, ilike, ilove, injoy, injoyed, injoying, injoys, intererested in, interested, interested in, interests include, interrested, intressted, likable, likd, like a lot, like everything about, like first one better, like it better, like it the way it is, like it's, like lots of, like the fact, like the first one better, like the first one the best, like the idea, like the idea of having, like the other one better, like this one best, like this one the best, like too, likeable, liked, liked the first one better, liked the other one better, likes, liket hat, liking, likings, lopve, love, love it, love that, love the, love the idea, love their, love this, loved, loves, loving, lust for, non annoying, not annoying, pleasantly supprised, pleasantly suprised, pleasantly surprised, prefer, prefered, prefers, proud, real interest, real interests, really like, really like what i do, really liked, really love, seem to like, seems to like, she like, she likes, stick to, stick with, sticks to, still like, still love, that like, they like, they love, thrilled, typical choice, unannoying, very much to my liking, we like, we love, well-liked, what i like most, what i liked best of all was, who like, will love