UNIVERSITY OF TARTU
DEPARTMENT OF ENGLISH STUDIES

SENTIMENT ANALYSIS OF SELECTED WORKS BY MARK TWAIN WITH THE STATISTICAL SOFTWARE *R*

BA thesis

ROLAND RAE

SUPERVISORS: Jane Klavan, PhD

Artjoms Šeļa, PhD

TARTU

2020

ABSTRACT

The capabilities of natural language processing have increased to the point where one could analyse large textual works as collections of data, enhancing potentially the analysis of works of fiction. The main purpose of such an analysis is to look for patterns in large texts without having to manually read every page. In the case of sentiment analysis, it provides a way of getting a rough estimate on the emotional disposition of a large textual data set. Sentiment analysis can be used to map out a curve of emotional disposition on a linear graph and thus portraying the general emotional highs and lows of a narrative. In addition, one could use the aggregated data to analyse word frequencies of several works of fiction in order to see if an author has any word choice preferences that are not apparent from manual reading.

The aim of this thesis is to explore the possibilities of sentiment analysis by focusing on a selection of literary works by Mark Twain such as: *The Adventures of Huckleberry Finn*, *The Adventures of Tom Sawyer*, *The Prince and the Pauper*, *Tom Sawyer, Abroad*, *Tom Sawyer, Detective*. The sentiment analysis is done in the programming language *R* and with the RStudio program, including software packages available through it. Since the process of doing sentiment analysis via RStudio is not fully automated, a script had to be written in order to accomplish the goal of doing sentiment analysis. The analysis of the selected works will be done in three ways. Firstly, by identifying the nine most frequent negative and positive sentiment words in every book separately. Secondly, visualizing the overall distribution of sentiment in every book in order to categorize the narratives into certain types of stories. Thirdly, by portraying the amount of negative and positive sentiment words that were used.

TABLE OF CONTENTS

INTRODUCTION 4

1 Literature review 6
- 1.1 Usage of sentiment lexicons 6
- 1.2 Sentiment Analysis in Literary and Cultural Research 9
- 1.3 Curves of narratives 14

2 Methodology 16
- 2.1 Overview of Methodology 16
- 2.2 The tidy text format 18

3 Sentiment analysis of Mark Twain's works of fiction 19
- 3.1 Most frequent positive and negative words 19
- 3.2 Sentiment curves per book 25
- 3.3 Amount of sentiment words per book 27

Discussion 28

CONCLUSION 33

LIST OF REFERENCES 35

Appendix 1A. The script for the RStudio environment 38
Appendix 1B. The data tables of the most used sentiment words 38
Appendix 1C. Data tables of curves of sentiment 38

RESÜMEE 39

INTRODUCTION

One of the main applications of digital humanities is to use computers to share and analyse textual data. One method for analysing textual data is using sentiment analysis, which is done by organizing textual input into machine readable tables and then cross referencing that data with sentiment lexicons. The output of such an endeavour is a rough estimate of the sentiment (emotional disposition) of the text. The aforementioned method has been used for both research and commercial purposes. The commercial use focuses on analysing customer feedback and comments in order to get a general outline of whether the clients are happy with a service or product. According to the Guardian: “sentiment analysis can help predict outcomes (from political events to product launches) and provide real-time insight into how people feel about a brand and its products – and how this sentiment compares with the brand's competitors” (Lawlor 2014. para 14). This is also the reason why most sentiment lexicons focus on whether a word or statement is of a positive or negative disposition. This thesis, however, will focus on the use of sentiment analysis for literary research.

The application of computational methods can yield new insights of textual data. Researchers can use sentiment analysis to process large textual data sets in order to get rough estimates of emotional disposition without the need to go through every page manually. This method allows researchers to plot sentiment words in a graph and portray estimates of emotional highs and lows of a story. Alternatively, one could also use sentiment data tables that were created during an analysis to analyse other aspects of a work of literature. For example, one could compare what kinds of sentiment words were the most frequent in an author's work. In addition, one could compare different works from

one author or analyse the works of multiple authors and compare the similarities and differences.

Text mining in humanities is a relatively new phenomena, which can be used in parallel with the important qualitative analyses done by the literary theorists. The statistical software *R* will be utilized for this thesis to analyse a selection of books by Mark Twain. The aim of this thesis is to use text mining techniques to analyse the sentiment words used in the selected works by Mark Twain in order to plot sentiment curves for every book and compare what kind of sentiment words were most frequently used in the works of fiction. The sentiment curves produced in this thesis are a rough estimate of the emotional flow of the books. The curves do not portray a specific emotion, but rather attempt to convey the shift from positive sentiment words to negative sentiment words through narrative time (beginning to end). The selected works are: *The Adventures of Huckleberry Finn*, *The Adventures of Tom Sawyer*, *The Prince and the Pauper*, *Tom Sawyer, Abroad*, *Tom Sawyer, Detective*. The criteria for the works of fiction were the following: it must be a single work of fiction and not a collection; it must be classic work of literature; it must be publicly available via the Gutenberg Library; it has to have text in one variety of English; all the works should be written by one author.

The works were acquired from the Gutenberg Library via a package in RStudio. RStudio is an integrated development environment (IDE) that provides a graphical interface for the programming language R. The analysis of the aforementioned works were done with RStudio and packages of software that are available through it such as `tidyverse`, `gutenbergr`, `ggplot2`, `dplyr`, `stringr`, and `tidytext`. The focus of the analysis will be on word sentiment, which will be portrayed in graphs. One graph will portray sentiment fluctuations within any one of the selected works of fiction and

several graphs portraying the positive and negative words that were used the most in the selected works individually.

1 Literature review

1.1 Usage of sentiment lexicons

Sentiment analysis is a text mining technique, i.e. the process of using specialized software to work through large amounts of digital data in order to find useful information. Text sentiment analysis is usually done with the aid of databases (sentiment lexicons) that are lists of sentiment words (opinion words) used in a language and their corresponding emotional connotations. Such connotations can have a binary value (positive vs. negative) or be descriptive of a mood (anger, sadness, joy, love, etc). The former can also have variations regarding sentiment intensity, which refers to the degree of a word's positive or negative sentiment (Liu 2015: 21). Liu (2015: 6) states that: "Sentiment words, also called opinion words, are words in a language that indicate desirable or undesirable states". Words such as *good*, *great*, and *beautiful* are positive sentiment words (Liu 2015: 6). A sentiment lexicon is a collection of sentiment words and phrases (Liu 2015: 10). In order to analyse the word sentiment values of a text in an automated way, as opposed to looking up the meanings of words manually with a dictionary, one needs to use a sentiment lexicon.

Sentiment can be divided into three categories: type, orientation, and intensity. There are several classifications for sentiment types such as linguistic-based, psychology-based, and consumer research-based. Sentiment intensity refers to the degree to which a word or phrase has a positive or negative sentiment. People tend to express the intensity of their emotions in two ways. One approach is to use words such as *good*, *wonderful*, and *amazing*, which all have a positive sentiment value, but differ in intensity.

The second way to express intensity of feelings is to use words that function as intensifiers or diminishers, which change the degree of sentiment. The former increases the intensity and the latter decreases it. Common intensifiers in the English language are words such as *extremely*, *very*, *dreadfully*, *really*, etc, while common dimishers are words such as *pretty*, *somewhat*, *barely*, *slightly*, etc. (Liu 2015: 20-21)

The sentiment properties to be analysed in this thesis will be sentiment orientation. Sentiment orientation can also be referred to as polarity, semantic orientation or valence (Liu 2015: 21). Sentiment orientation can either be positive, negative or neutral (Liu 2015: 21). The data that was gathered from the selected works of Mark Twain and analysed with the help of RStudio will represent the sentiment values of words in graphs as either positive (+1) or negative (-1). This thesis will focus on the semantic orientation of individual words and not on the sentiment of idiomatic constructions, phrases or sentences. Sentiment analysis is an imperfect method, but it is still used because it provides a proxy for sentiment dynamics within text. According to Liu (2015: 11) a sentence may not express sentiment despite the fact that it has a sentiment word in it. The two main types of sentences where this occurs are interrogative and conditional sentences (Liu 2015: 11). For example, the sentence "Could you tell me, which iPhone is good?" has a positive sentiment word in it but does not express a positive opinion regarding an iPhone. The same goes for the conditional construction "If I find a good computer in the store, then I will purchase it", which does not really express a positive or negative sentiment.

Another issue in sentiment analysis is the phenomenon known as sarcasm (Liu 2015: 11). Sarcasm refers to a form of writing or speech that has an opposite meaning compared to the words that were used (Liu 2015: 82). Sarcasm complicates sentiment analysis, because when one writes or says something negative, then the meaning is positive

and vice versa (Liu 2015: 82). Liu (2015: 82) claims that there have been initial attempts of dealing with sarcasm, but states that "/.../ our knowledge about it is still very limited." There are also sentences that have a sentiment value, but do not have a sentiment word in them. For example, "This television uses a lot of electricity", which implies a negative opinion about a device, because it consumes a large amount of a resource, but the sentence does not have a negative sentiment word in it. A similar issue for sentiment analysis is the metaphorical use of words in a sentence (Liu 2015: 11).

In lexicon based approaches words in an analysed text are matched with words in a dictionary and are assigned corresponding orientation values. The sentiment orientation values of all sentiment expressions in a document are summed up in order to classify a document as either positive, negative or neutral (Liu 2015: 59). All the analysis of the selected works of Mark Twain used a sentiment lexicon that was created by Minqin Hu and Bing Liu (2004). One aspect of the analysis was to create a graph with the relative sentiment orientation values for all the books in order to evaluate which book had the most positive or negative sentiment orientation based on word sentiment without negation or other sentiment shifting or reversing factors. In addition to sentiment shifting factors, there are also irrealis moods, which usually indicate nonfactual contexts (Liu 2015: 61). Irrealis moods are markers such as conditional markers (*if*), modals, negative polarity items (*any, anything*), certain verbs (*expect, doubt*), questions, and words enclosed in quotation marks. Irrealis blocking is a strategy used to ignore such markers (Liu 2015: 61). According to Liu (2015: 61) this strategy is used, because "/.../ it is hard to reliably determine when such sentences express sentiment and when they do not. They are thus ignored."

1.2 Sentiment Analysis in Literary and Cultural Research

The main purpose of sentiment analysis is to extract opinions or moods from textual documents with computational methods in order to gain further insights. Sentiment analysis has been used on works of fiction by organizations such as the Lord of the Rings Project (Johansson 2012), Eindhoven University of Technology (Morin et al 2017), Stanford University (Heuser et al 2016), University of Vermont (Reagan et al 2016), Washington State University (Jockers 2015), and the University of Cambridge (Brand et al 2019).

Sentiment analysis can be used to find all kinds of interesting information regarding works of literature. For example, Heuser et al. (2016) used sentiment analysis and crowdsourcing to create an "emotional geography" of London, which is based on works of fiction from a corpus of roughly 5,000 novels published between the years 1700 and 1900. The program that was used for the analysis was developed by the Stanford Linguistics department and had about 1,700 negative and 1,300 positive terms in its sentiment lexicon. The Stanford program was trained on a corpus of Wall Street Journal articles and similar sources (Heuser et al. 2016). While historical articles from the Wall Street Journal may be helpful in creating a corpus of the general vocabulary of a time period, it may also cause some inaccuracies in the corpus, because news articles and literature have a different narrative structure and feature different word usage. The graphs that resulted from the aforementioned research portray what kind of sentiment, positive (happy) or negative (frightening), was used when authors of the selected 5,000 works of fiction referred to certain parts of London. Heuster (et al 2016: 10) claimed that:

> "/.../ we found empirical evidence that supported existing theories about emotions in public; we showed how established narratological polarities (foreground/background, story/discourse) preside, not only over the temporality of narrative, but over its geography as

> well; and we discovered striking discrepancy between real and fictional geography, while also sketching the first lineaments for a future 'semantics of space'."

Sentiment analysis can also be used on other media in order to research culture. Such an endeavour was undertaken by Charlotte O. Brand, Alberto Acerbi, and Alex Mesoudi (2019). Their research used 50 years (1965–2015) of song lyrics in order to estimate long-term cultural evolutionary dynamics. Brand et al.'s (2019) preliminary analysis discovered an increase in the use of emotionally negative words, while words relating to a positive emotion decreased. An example of this is the infrequency of the word *love*, the use of which halved over the time period that was analysed and the term *hate* became substantially more frequent (Brand et al. 2019).

Brand et al. (2019) factored in biases such as success, prestige, content and the emotional trends that derive from success bias, which comes from artists copying best-selling songs from the previous year. Prestige bias refers to the emotional trends that derive from mimicking prestigious songs from the preceding year by other artists (Brand et al. 2019). The general psychological preference towards lyrics that have an emotionally negative meaning ranking higher in charts is referred to as content bias (Brand et al 2019).

Brand (et al 2019: 4) used a cluster function in R (`refinr`) to merge and cluster artists' names that referred to the same artist while having slightly different spelling such as capitalised or non-capitalised first word. When researching sentiment analysis, the program R is mentioned and used often, indicating that it is a powerful tool for mining and analysing data. Brand et al. (2019: 5) used R version 3.6.0 and a package called `rethinking`. Brand et al. (2019: 9-10) discovered that the billboard dataset containing the top-100 songs from the years 1965 to 2015 revealed a content bias in transmission of

negative lyrics and unbiased transmission of positive lyrics. Brand et al. (2019: 10) states that:

> "The effect of unbiased transmission is, however, the largest and most consistent in all of our models. This result suggests there may be an effect of random drift, or random copying, in the emotional content of song lyrics over time. /.../ Thus, rather than song-writers being influenced by the most prestigious or successful artists, they may simply be influenced by the emotional content of any of the available song lyrics in the previous timestep, which may happen to increase in negativity or decrease in positivity owing to small fluctuations."

In addition to the previously mentioned research, sentiment analysis has also been used to research reviews regarding works of literature on Goodreads. Parksepp's (2019: 2) research used the program SentiStrength to analyse Goodreads reviews of 10 classic American works of fiction in order to evaluate how much sentiment is expressed in one star reviews compared to five star reviews and which of the selected books had the highest sentiment in its review. According to Parksepp (2019: 13), SentiStrength is freely available for academic purposes and compatible with both Microsoft Windows and MacOS. Professor Thelwall from the University of Wolverhampton created the SentiStrength program, with the aim of analysing short informal texts (Parksepp 2019: 15). Parksepp (2019: 13) copied 20 positive and 20 negative reviews per book from Goodreads.

Parksepp (2019: 13-14) modified the paragraph breaks for the reviews in order to get one sentiment score per review. In addition, quotes from the books that were in the reviews were removed with the aim of cleaning up the data (Parksepp 2019: 14). SentiStrength uses a scale of 1 to 5 for both negative and positive sentiment words (SentiStrength n.d: About). The aforementioned program provides a positive and negative score for each sentence, which means that every sentence has two scores one negative and the other positive (Parksepp 2019: 14).

Parksepp's empirical analysis focused on the Goodreads reviews of books such as Ralph Ellison's *Invisible Man*, Edgar Allan Poe's *The Complete Stories and Poems*, Ernest

Hemingway's *The Old Man and the Sea*, Vladimir Nabokov's *Lolita*, Walt Whitman's *Leaves of Grass*, Herman Melville's *Moby-Dick; or, the Whale*, Toni Morrison's *Beloved*, Mark Twain's *The Adventures of Huckleberry Finn*, William Faulkner's *Absalom, Absalom!*, and *The Sound of the Fury* (Parksepp 2019: 16-17). Parksepp (2019: 18) picked the reviews based on the support they gained from the Goodreads community (in the form of likes), because she felt that this method would yield interesting correlations with the sentiment data. The book *Lolita* had the highest positive sentiment of 4.05 and highest negative sentiment of -4.3 for 5 star reviews (Parksepp 2019: 18). The book *Leaves of Grass* had the lowest negative sentiment score of -2.5 for five star reviews (Parksepp 2019: 18). Parksepp (2019: 18) discovered that the five star reviews had the highest average sentiment for all books. The sentiment scores for *The Adventures of Huckleberry Finn* were: 3.45 (positive sentiment) and -3.6 (negative sentiment) for five star reviews, 2.55 and -3.55 for one star reviews, and an average score of 3.3 (Parksepp 2019: 19). Parksepp (2019: 19) claimed that:

> "So it seems that reviewers, who write 5 star reviews, do so with more emotion than those who write 1 star reviews. While many of the negative reviews use pungent language, their reviews are not long enough to have that emotion registered as very strong."

There have been other researchers who have used sentiment analysis for graphically portraying how sentiment words are used in a work of fiction. While this thesis focuses on a few selected works by Mark Twain, the Lord of the Rings Project analysed a bigger data set and graphically displayed the result of their research in a similar way to the work of this thesis. Johansson (2012) from the Lord of the Rings Project analysed the works of John Ronald Reuel Tolkien and published the results of the distribution of sentiment words as linear graphs. This method is very effective in displaying the mood of works of fiction throughout the narratives. Johansson (2012) analysed the sentiment of every sentence and

then averaged it over each page, thus providing a neat graph where the observer can keep track of what page and chapter the sentiment highs and lows were found in. The software that was used for the aforementioned research was an application programming interface (API); an API is "a way of communicating with a particular computer program or internet service" (Cambridge Dictionary 2020). The API used for the analysis of Tolkien's works is called Sentiment140, which was created by Stanford Computer Science graduate students such as Alec Go, Richa Bhayani, and Lei Huang (Johansson 2020).

It can be seen from the overview above that sentiment lexicons have been utilized for analysing a variety of different textual data such as comments on Goodreads, song lyrics and literature. The purpose of such endeavours is to create a rough estimate of the emotional disposition of a text. The work of Parksepp (2019) attempted to evaluate the emotions of Goodreads' comments with computational methods rather than close reading. Other researchers focused on analysing works of literature or song lyrics in order to find how the sentiments within such works express or correlate to people's emotions. Brand et al. (2019) focused on how the sentiment words within song lyrics could be used to describe cultural trends such as shifts towards negativity and how a song from previous years may influence songwriters. While the aforementioned research used song lyrics, Heuser et al. (2016) used works of literature to analyse what kind of emotions were ascribed to parts of historical London. The emotions depicted in a work of fiction can also be depicted as a linear graph in order to see the general outline of where sentiment words form clusters and whether they have a positive or negative connotation. Johansson (2012) analysed works of fiction by Tolkien and produced linear graphs that provide an estimate on how sentiment words are distributed in the text. Johansson's (2012) work is similar to the research done in this thesis in two ways. This thesis will first focus on one author and secondly, it will

utilize sentiment lexicons in order to plot an estimate of the changes in sentiment through narrative time on a linear graph.

1.3 Curves of narratives

Analysing sentiment through narrative time enables a researcher to find a rough estimate of the change in the sentiment that is being expressed. It is possible to plot the use of sentiment words in a graph in order to portray a linear estimate of the emotional highs and lows of a story. The utilization of the Gutenberg Library is ideal for such endeavours, because the works in it are available for free and have been digitized.

The notion of analysing works of fiction with computerized methods has been around for at least a decade. Kurt Vonnegut (Flood 2016. para 2) claimed that "There is no reason why the simple shapes of stories can't be fed into computers; they are beautiful shapes". This statement was considered a challenge by Matthew L. Jockers (2015), who used sentiment analysis to analyse a corpus of 50,000 novels. Jockers (2015) used the program R to analyse works of fiction and went further by creating his own R package called `syuzhet`. During his research, Jockers (2015) discovered that sentiment fluctuations can function as a proxy for plot movement. Jockers (2015) borrows a distinction between fabula and syuzhet from Russian formalists, the former being a general disposition of events that make a story and the latter a progression of a narrative from beginning to end as represented in a text by an author. This is important to understand the focus of sentiment analysis, which may not portray an important moment in the story as a spike in a graph, but rather portrays what parts of a book had the biggest clusters of sentiment. In other words, the focus is not on the order of the events, but rather on how an author presents events to his or her readers (Jockers 2015).

Similarly to the work of Jockers (2015), Reagan et al. (2016) also used sentiment analysis to plot the distribution of sentiment words in a linear graph. Reagan et al. (2016: 1) selected 1,327 works of fiction from the Gutenberg Library with the aim of analysing the emotional arcs of the books. Reagan et al. (2016: 1) claimed that stories tend to form patterns and emotional trajectories that are meaningful to people. A set of six core arcs were found, which are the "essential building blocks of complex emotional trajectories" (Reagan et al 2016: 1). One of the books that Reagan et al. (2016: 3) analysed was JK Rowling's *Harry Potter and the Deathly Hallows* and stated that despite the complicated plot of the book, the emotional arc of their graph portrayed each sub-narrative. The selection of the books involved criteria such as it must be in English; it must have 20,000 - 100,000 words; it must have more than 40 downloads from the Project Gutenberg; and it must correspond with the Library of Congress Class of English fiction (Reagan et al. 2016: 4). Works that were labelled as collections or poems were excluded from the research (Reagan et al. 2016: 4).

The researchers claimed that there are six emotional arcs: rags to riches, tragedy, 'Man in a hole', Icarus, Cinderella, Oedipus (Reagan et al. 2016: 5). For example, the Cinderella shape of a story is characteristic for its rise-fall-rise pattern of sentiment distribution. Interestingly, books that have an emotional arc such as Icarus, Oedipus or 'Man in a hole' had the most downloads (from Project Gutenberg) and may be considered the most successful (Reagan et al. 2016: 10). Reagan et al (2016: 11) concluded their paper by stating that:

> "We are producing data at an ever increasing rate, including rich sources of stories written to entertain and share knowledge, from books to television series to news. Of profound scientific interest will be the degree to which we can eventually understand the full landscape of human stories, and data driven approaches will play a crucial role."

The analysis of curves of stories could yield new insights into both old and new works of fiction. Thus, it is worth investigating what kinds of possibilities are currently available in order to gain further knowledge about nuances in literary narratives. This thesis will attempt to map the sentiment of a few of Mark Twain's books in order to visually portray the fluctuations of positive and negative sentiment. The aforementioned goal was achieved by the use of the *R* programming language (The R Foundation 2020), RStudio (2020), the tidy text format (Silge et al. 2019) and a sentiment lexicon created by Hu and Liu (2004).The following sections will elaborate the methodology and portray the results of the work of this thesis.

2 Methodology

2.1 Overview of Methodology

The initial part of the analysis is the acquisition and cleaning of the text. The books were acquired from the Gutenberg Library via an *R* package called `gutenbergr`. The cleaning of the data was done by organizing the words of the selected books into tables and removing grammatical words (stop words). Organizing the text into tables creates a neat format that is machine readable. One such approach is the application of the tidy text format, which provides guidelines and a general structure for how one should arrange a dataset. Once the data has been cleaned it is important, for sentiment analysis, to initiate the analysis of the tidy data by cross referencing the words from a source material to a sentiment lexicon. Thus, creating new data tables that can be used for a graphical output such as bar charts and linear graphs. The final part of analysing textual data with computational methods is the creation of graphical outputs in order to portray what kinds of characteristics a text has.

An important aspect of doing sentiment analysis is selecting a sentiment lexicon. For the RStudio environment the options were rather limited. Three lexicons were designed to work within the R based environment, but unfortunately only one of them was updated to work with the most current version of R. Thus, the sentiment lexicon created by Hu and Liu (2004) was the only viable option for doing the analysis in R.

In order to analyse several books a script had to be written to accomplish the analysis in the RStudio environment. The entire script with comments regarding the commands, which is 358 lines long, can be found in the software development platform GitHub. The link to the script is provided in Appendix 1A. The Script is divided into four parts: data cleaning and processing, analysis of most frequent words, and analysis of the emotional curves of the narratives, and an estimate of the total amount of positive and negative sentiment words per book. The removal of grammar words (stop words) was necessary, because negation and sentence structure were not part of the aim of this thesis.

In order to portray the most frequently used sentiment words all the words that were turned into tokens were compared to the sentiment lexicon and counted for frequency by additional commands. The results of the sentiment curves were achieved by a similar process. The main differences were in how the program counts the words and what kind of graphical output format was used. In the analysis of sentiment curves, words were grouped into 1,500 word clusters in order to make the graphical output more clear. Otherwise the graph would look very chaotic and it would be difficult to see a general outline for the books. The graphical output of the sentiment curves is done via a linear graph in order to portray the general outline of sentiment values from the beginning of a book to the end.

The script was designed with the aim that other researchers, who are interested in checking the results of this thesis or replicate the process on another set of literary works,

are able to understand what the lines of code actually do and thus there are a lot of comments inside the script that explain the commands. In addition to the comments, some technical knowledge is required to produce and export the graphs. For example, running the script from start to finish will produce all the relevant data in the RStudio data environment, but in order to see individual graphical outputs one needs to run the plot commands one by one. There are no commands that export the graphical output of the visual data, because RStudio has a function to export visual data via a button, which is aptly named *Export*, thus making the creation of any additional lines of code with a similar function unnecessary. In the event that one wishes to export data tables, the final commented lines of the script feature a method for data table extraction. Alternatively, one could download the tables used in the present thesis from the links provided in Appendix 1B and Appendix 1C.

2.2 The tidy text format

The tidy text approach requires one to tokenize text and rearrange it into a table. The goal is to use the tidy text format, which is "a table with one-token-per-row" (Silge 2019. para 2. Chapter 1). A token is described as a unit of text, which may be a sentence, word, n-gram, or paragraph depending on the kind of analysis that is being done (Silge. 2019. para 2. Chapter 1). The script that was written for this thesis counts a single word as a token and cross references the tokens to the sentiment database created by Hu and Liu (2004) in order to create a large table of data that portrays a rough estimate of what kind of sentiment is expressed in the selected works by Mark Twain. Hu and Liu (2004) created the sentiment database (lexicon) for commercial purposes, mainly for the analysis of contemporary sources of text. The aforementioned lexicon is the most up to date database regarding sentiment words that can be used via the packages in RStudio. Ideally, the

sentiment analysis should utilize a sentiment lexicon that was created with historical sources similar in date to the original text, because some words have changed in meaning and can thus cause some irregularities in the data.

In addition to the use of tokens, the data needs to be organized into orderly tables and be comprehensible for computational commands. According to Wickham (2014: 1), data preparation is an important step in the course of an analysis and must be done with great attention to detail. The standard of tidy data created to "facilitate initial exploration" and "simplify the development of data analysis" (Wickham. 2014. 1). The tidy data format makes it easy for a computer or analyst to access the necessary information (Wickham 2014: 5). The approach of tokenizing the textual data and organizing it into orderly tables is helpful for the program to read and compute the sentiment values of the tokens, but also for anyone who wishes to manually double check the data.

It is important to consider the characteristics of a text and how that might impact the analysis. For example, the book *Adventures of Huckleberry Finn* was written as mostly dialogue, which would complicate an analysis of sentence structures because dialogue may not be as eloquent and grammatically correct as the sentence structures of a third person narrator. This thesis focuses on the sentiment of words individually and thus complications regarding grammatical cohesion do not apply. However, it should be noted that elements of dialogue such as colloquialism and purposefully misspelled words may not register as units of sentiment because there may not be an entry for them in the sentiment lexicon.

3 Sentiment analysis of Mark Twain's works of fiction

3.1 Most frequent positive and negative words

The aim of this analysis is to portray which sentiment words were used most frequently in the books that were selected for this thesis. The y-axis of the graphs displays the sentiment tokens, while the x-axis represents the number of times they occur within a given book. Figure 1 depicts the nine most used sentiment words in the book *The Adventures of Huckleberry Finn*. The negative sentiment words in Figure 1 with their respective frequencies are: *trouble* (84), *miss* (78), *dead* (76), *dark* (70), *poor* (51), *struck* (50), *bad* (41), *blame* (39), *fool* (38). While the nine most used positive sentiment words in Figure 1 are: *pretty* (162), *mighty* (71), *easy* (51), *glad* (46), *ready* (42), *free* (41), *luck* (30), *comfortable* (30), *safe* (28).

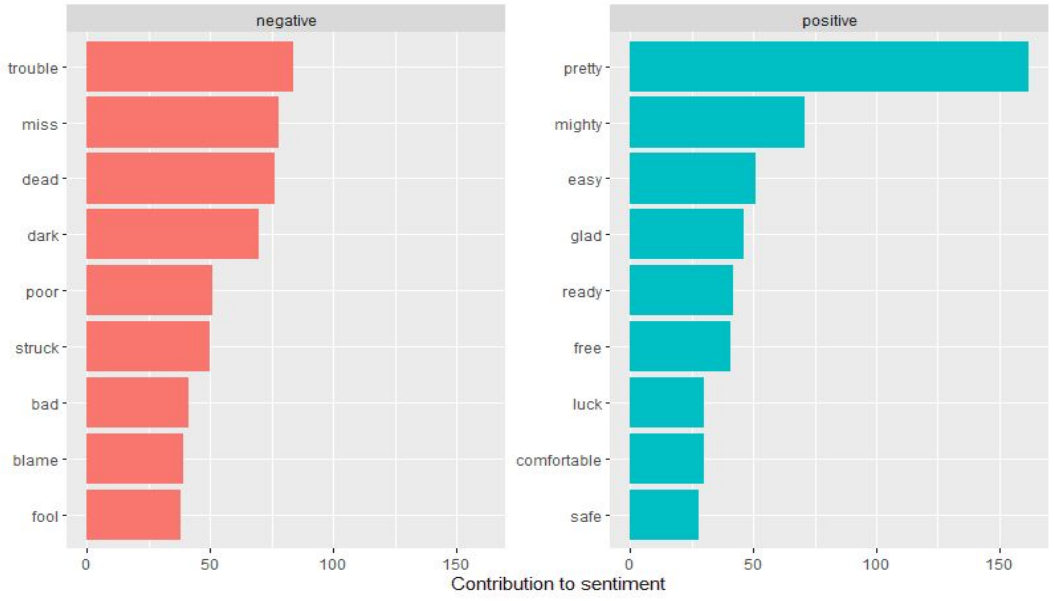

Figure 1. Top 9 negative and positive sentiment words in *The Adventures of Huckleberry Finn*

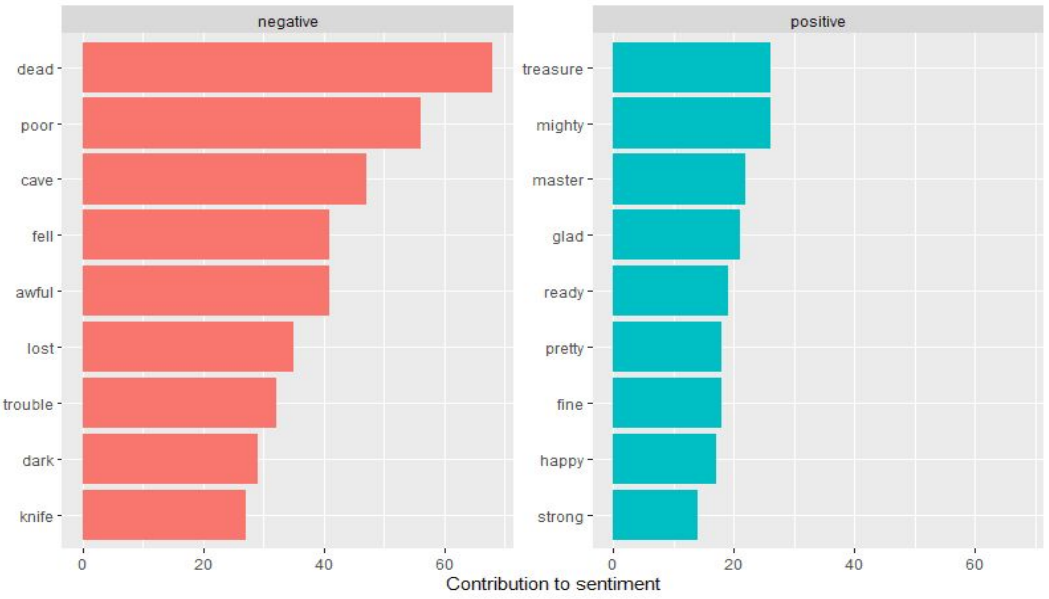

Figure 2. Top 9 negative and positive sentiment words in *The Adventures of Tom Sawyer*

Figure 2 depicts the most frequent sentiment words in the book *The Adventures of Tom Sawyer*. The most frequent positive words are: *mighty* (26), *treasure* (26), *master* (22), *glad* (21), *ready* (19), *fine* (18), *happy* (17), *strong* (14); while the most frequent negative words are: *dead* (68), *poor* (56), *cave* (47), *awful* (41), *fell* (41), *lost* (35), *trouble* (32), *dark* (29), *knife* (24).

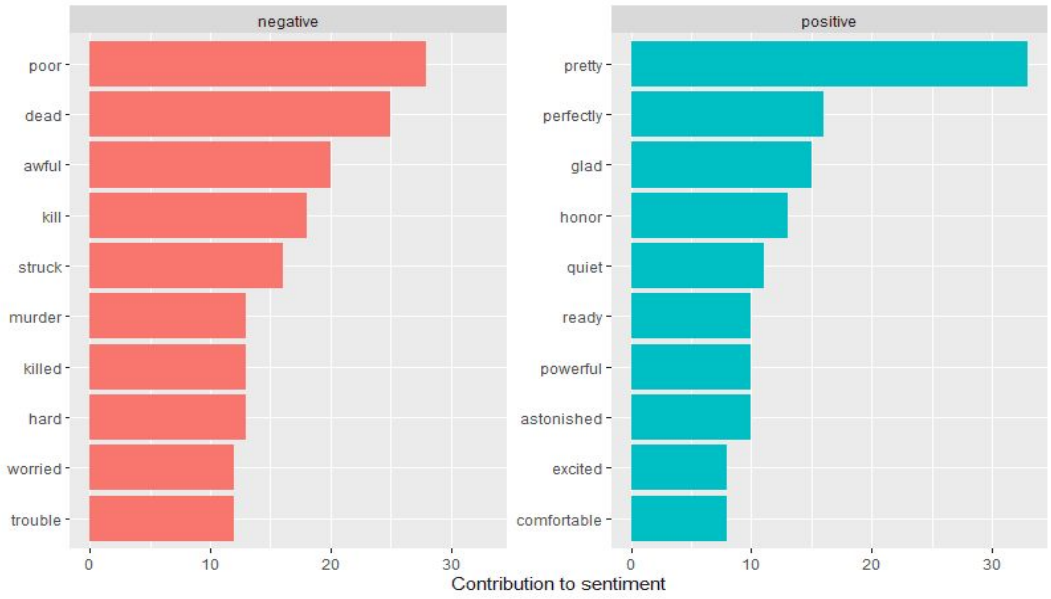

Figure 3. Top 9 negative and positive sentiment words in *Tom Sawyer, Detective*

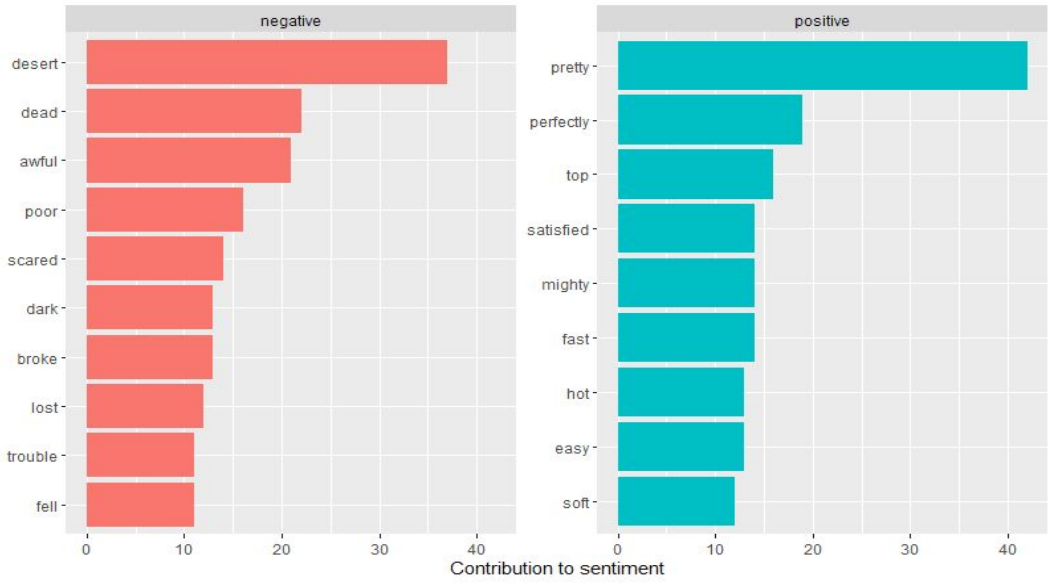

Figure 4. Top 9 negative and positive sentiment words in *Tom Sawyer Abroad*

The book with the most uses of the word *dead* is *The Adventures of Huckleberry Finn* with 76 occurrences (Figure 1). The aforementioned word was in the top nine of all the books. The book with the least references to the word *dead* (22) is *Tom Sawyer Abroad* (Figure 4). In addition, the word *poor* was also in the top nine charts of all the books used in this analysis. The book with the most references to *poor* is *The Prince and Pauper* with 108 occurrences (Figure 5), while *Tom Sawyer Abroad* has merely 16 (Figure 4). The only book to frequently use the words *kill* (18), *murder* (13), and *killed* (13) is *Tom Sawyer, Detective* (Figure 3). Cross referencing the books *The Prince and the Pauper* (Figure 5) and *The Adventures of Tom Sawyer* (Figure 1) showed that the word *lost* appeared 35 times in both books and merely 12 times in *Tom Sawyer Abroad* (Figure 5). The word *mighty* appeared among the top nine sentiment words in three books. The frequencies of the aforementioned word are as follows: 14 times in *Tom Sawyer Abroad* (Figure 4), 26 times in *The Adventures of Tom Sawyer* (Figure 2), 71 times in *The Adventures of Huckleberry*

Finn (Figure 1). The word *master* is used 23 times and is marked as a positive sentiment word in the book The Prince and Pauper.

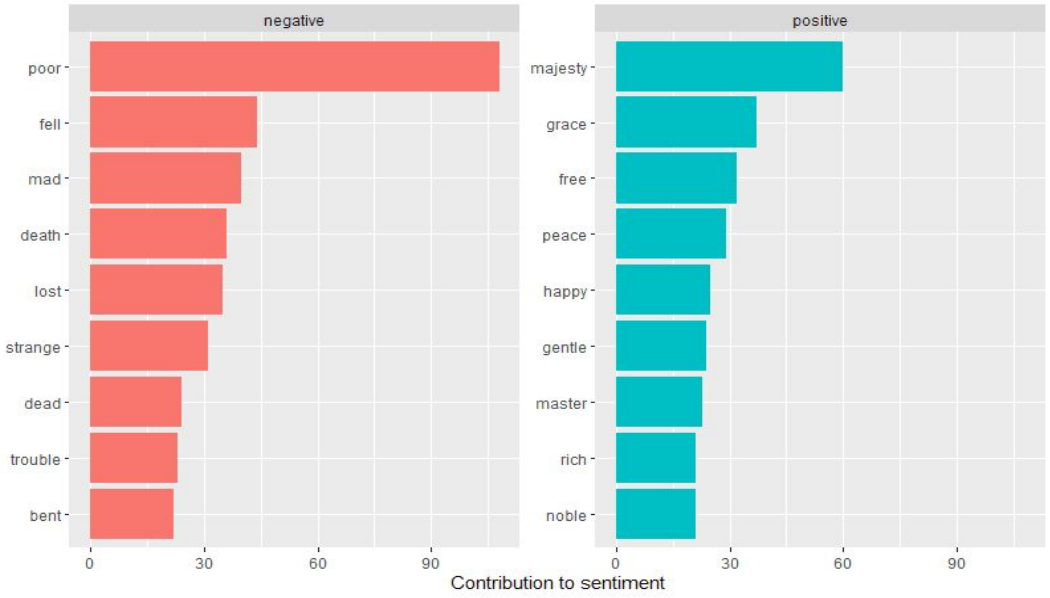

Figure 5. Top 9 negative and positive sentiment words in *The Prince and the Pauper*

The word *pretty* seems to be very frequently used throughout the selected works by Mark Twain. The aforementioned word is the most frequent positive sentiment word in three of the five selected books. The only book where it is not in the top nine chart is *The Prince and the Pauper* (Figure 5). The list of most frequent positive sentiment words is provided in Table 1. In the book *The Adventures of Huckleberry Finn* the word *pretty* (162) is almost twice as frequent as the most negative word *trouble* (84); see Table 2 for the most frequent negative sentiment words. The complete data tables with lists of the sentiment words with their respective frequencies can be accessed via Appendix 1B.

The relative frequencies in Tables 1 and 2 were calculated by dividing the number of occurrences of a word with the total number of sentiment words of the same book and then multiplying the result by a 100. The aforementioned equation provides a percentage that portrays how frequently a certain sentiment word was used out of all sentiment words.

Table 1. Most frequent positive sentiment words per book.

Names of the books	**Word(s)**	**Number of occurrences**	**Total number of sentiment words**	**Relative frequency**
Tom Sawyer Abroad	pretty	42	474	8.86 %
Tom Sawyer, Detective	pretty	33	400	8.25 %
The Adventures of Huckleberry Finn	pretty	162	794	20.66 %
The Adventures of Tom Sawyer	mighty, treasure	26	1347	1.93 %
The Prince and the Pauper	majesty	60	1456	4.12 %

Table 2. Most frequent negative sentiment words per book.

Names of the books	**Word**	**Number of occurrences**	**Total number of Sentiment words**	**Relative frequency**
Tom Sawyer Abroad	desert	37	474	7.81 %
Tom Sawyer, Detective	poor	28	400	7 %
The Adventures of Huckleberry Finn	trouble	84	794	10.58 %
The Adventures of Tom Sawyer	dead	68	1347	5.05 %
The Prince and the Pauper	poor	108	1456	7.42 %

The word pretty in *The Adventures of Huckleberry Finn* is roughly twice as frequent (Table 1) as the most frequent negative sentiment word trouble (Table 2). It is interesting to note that while the most frequently used positive words in the first three books in Tables 1 and 2 are more frequent than their negative counterparts, the last two books in the aforementioned tables show that the most frequent negative sentiment words are much more frequent than the most frequent positive sentiment words.

3.2 Sentiment curves per book

Reagan et al. (2016) state in their research that there are six basic shapes of stories. These shapes will be compared to the sentiment analysis graphs that were produced for this thesis with the aim of discovering the patterns or shapes of the selected works by Mark Twain. The sentiment analysis of the selected works by Mark Twain produced linear graphs that denote the highs and lows of sentiment throughout the selected books. Figure 6 depicts the general outline of sentiment fluctuations.

The y-axis of Figure 6 represents the change in sentiment throughout the story and the x-axis represent sequences of words (chunks) that were counted from the beginning to the end of every book. In order to see a clear pattern, an appropriate number of words needed to be counted as one chunk in order to portray the shift in sentiment throughout the stories. One chunk is equal to the average sentiment value of 1,500 words. It should also be noted that using the same chunk size for all the books produced slightly different curves due to the difference in the length of the stories. In Figure 6 the shorter books *Tom Sawyer, Abroad* and *Tom Sawyer, Detective* have much smoother lines than those of the larger books.

The chunk size for the analysis of sentiment curves (shapes of stories) was chosen to be 1,500 words for two reasons. One, it works well as an average value in order to portray all of the books in one graph despite the differences in length. Two, it distorts the sentiment lines just enough to portray the aforementioned issue of choosing an optimal chunk size when doing sentiment analysis

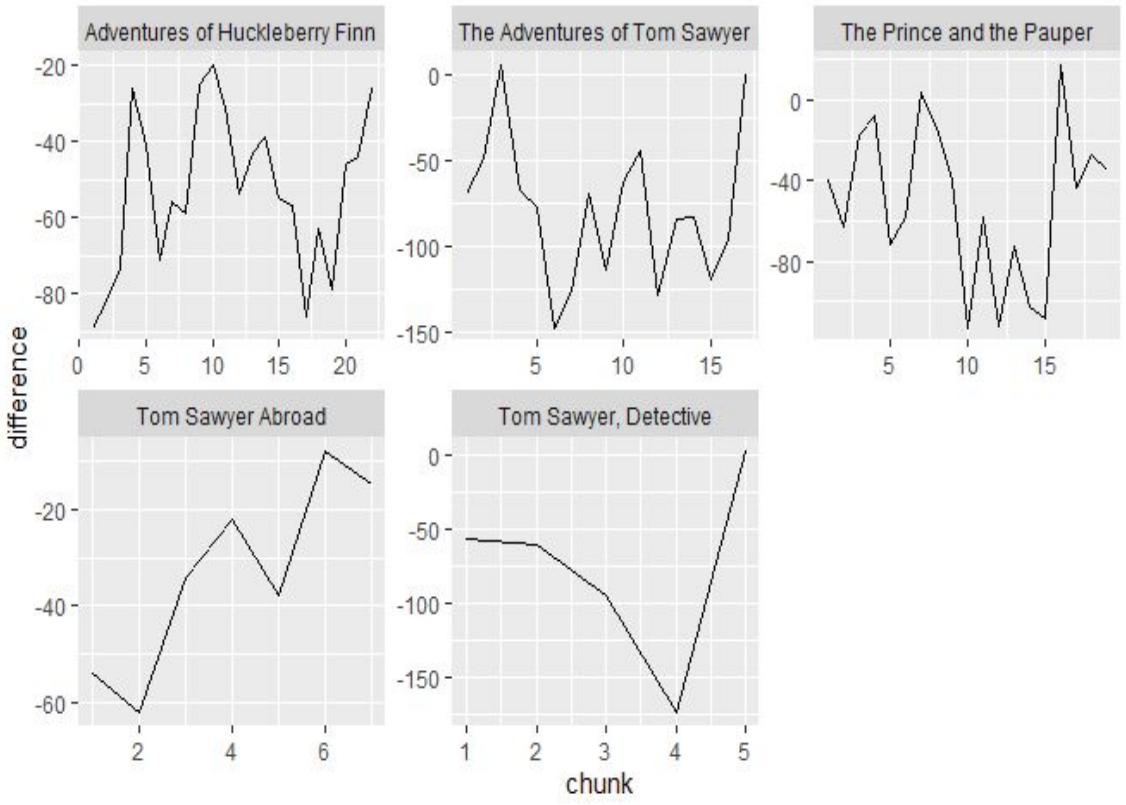

Figure 6. Linear sentiment graphs with a chunk size of 1500 words

Figure 6 shows that the book *Tom Sawyer Abroad* has a sentiment curve that depicts a mostly steady rise and thus can be considered to have the shape of a "rags to riches" (rise in positive sentiment) story. The book *Tom Sawyer, Detective* has the shape of a "man in a hole" story (negative to positive), because the sentiment steadily decreases only to rapidly increase towards the end of the story. A "Cinderella" pattern (increase - decrease - increase) can be seen in the sentiment analysis of *The Prince and the Pauper*, which has an increase in positive sentiment that is followed by a decline towards negative sentiment, which is followed by an increase towards positive sentiment, that might indicate a happy ending. A similar pattern is seen in the sentiment analysis of *The Adventures of Tom Sawyer*, which also initially rises in positive sentiment, then falls towards negative sentiment only to be followed by a rise in positive sentiment. The book *Adventures of*

Huckleberry Finn also fits the description of a "Cinderella" story with its initial increase in positive sentiment that is followed by both a decrease and subsequent increase in sentiment. All the books used for this analysis end on positive increase in sentiment compared to the initial starting position of the narratives. The most frequent shape of stories was "Cinderella", which was used for three out of the five books.

3.3 Amount of sentiment words per book

The total amount of sentiment words is portrayed in Figure 7. The y-axis represents the percentage of sentiment words out of the sum of all words that were used in the books. Columns pointing downward on the left-hand side of a book title represent the percentage of negative sentiment words used in a book. Inversely, the columns pointing upward on the right-hand side of a book title represent the percentage of positive sentiment words. The book *Tom Sawyer, Detective* has the largest percentage of negative sentiment words used out of all the five books that were analysed (Figure 7). *The Prince and the Pauper* has the largest percentage of positive sentiment words used compared to the other books in the analysis. The smallest percentage of negative sentiment words was used in *Tom Sawyer Abroad*. The book *Adventures of Huckleberry Finn* had the smallest percentage of positive words out of all the five books that were analysed. A large percentage of negative sentiment words might indicate how much conflict or tragedy is depicted in a story.

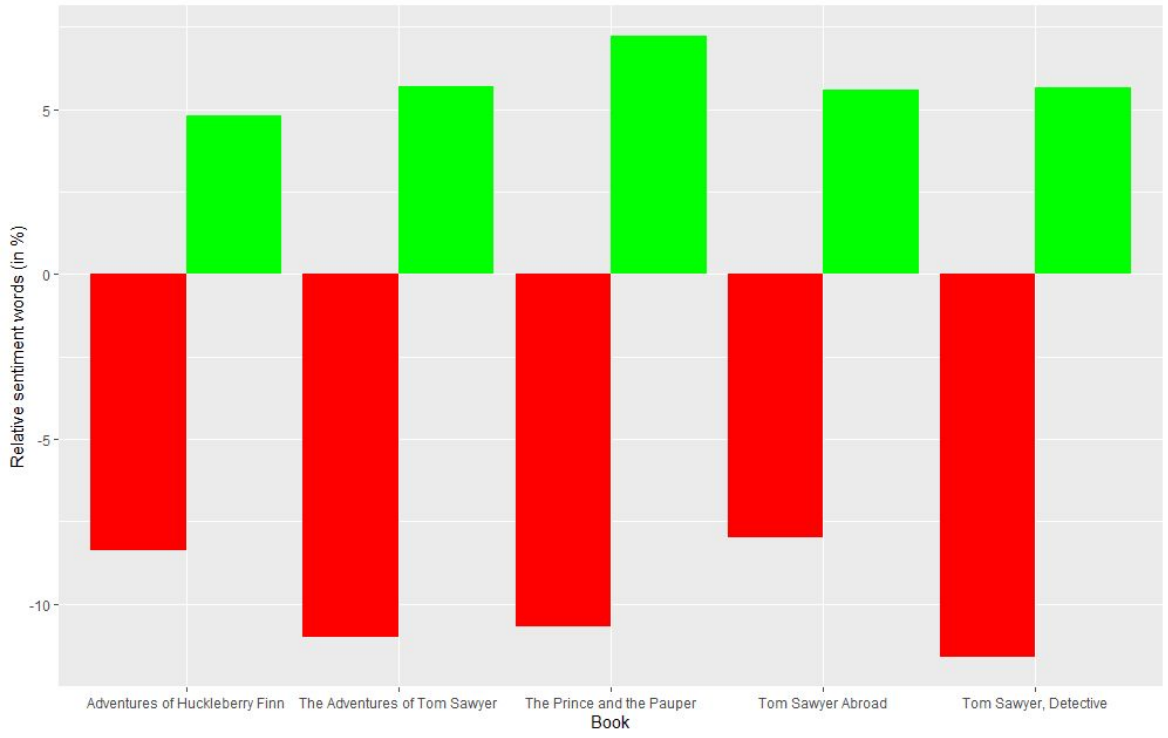

Figure 7. Percentage of sentiment words

Discussion

Sentiment analysis can yield new insights into works of fiction but should be used with a great deal of caution, because there are several linguistic aspects that need to be taken into consideration in order to extract meaning from works of literature with statistical software. Sentence structure can alter the meaning of a sentiment word and thus mislead a program. The same can be said for word constructions that have negation or intensifiers, which in the case of the former flip the sentiment value and in the latter increase it.

One should also consider the historical context of the work that is analysed. If a work of fiction is analysed with a sentiment lexicon that is created based on contemporary sources, then the data needs to be meticulously checked via manual reading. There are, however, complications when choosing a sentiment lexicon. The Stanford University lexicon and analysis tools are written in the programming language Java and are thus

difficult to tie to an analysis done in RStudio. Thus, the Stanford University lexicon was not used to analyse the selected works by Mark Twain. Ideally one could apply both lexcions and analyse the differences in data output in order to gain insights of how the choice of lexicon affects research results. Future researchers should keep in mind what programming environment they are capable of working with, which lexicons are available for that environment, and what consequences that can have to their research.

Sentiment analysis could be a useful tool to enhance literary research, because it provides a method to work with large textual datasets. Such an approach could aid researchers compare different works of text to each other or even different authors in order to find more insights into what makes them different. It would be ideal, of course, to utilize sentiment lexicons that are historically appropriate for a selected text, but since this is a rather new field with a steep learning curve, that might not always be possible.

One example of a problematic sentiment word from this analysis is the word *master*, which was among the most frequently occurring sentiment words in the book *The Adventures of Tom Sawyer*. It is important to keep in mind the historical context of a story (setting) where the practice of owning slaves may have been in effect. Hence, it is difficult to claim that the word *master* has a positive or negative sentiment, because it very much depends on the way it is used within a story. If *master* is used to refer to someone being skilled in their profession, then it is positive, but if it is a reference to someone owning people, then it is negative.

The word *master* was used 22 times, in the book *The Adventures of Tom Sawyer*. In order to accurately state the sentiment value, one would have to manually go over every reference and mark it appropriately. The aforementioned book uses the word *master* to refer to the concept of owning people “call any being master or obey anybody”, but for the

most part the word *master* is used as a reference towards a schoolmaster (Twain 2006b). The book *Adventures of Huckleberry Finn* makes a clearer reference to the practice of slave owning in the line "and if their master wouldn't sell them [people]" (Twain 2006c). It should be noted that the setting of the book *The Prince and the Pauper* is not in slave owning American South and thus, the word *master* can be viewed with a positive connotation. Here are a few examples of positive uses of the word master from The Prince and the Pauper such as *Master of the Wardrobe, master thy tongue, O most noble master* (Twain 2006a). When analysing historical texts it is likely that sentiment values of certain words will be determined not merely by grammatical structure, but also temporal context and it is thus a problematic aspect for distant reading.

It is very rewarding to see patterns in the data that at the very first glance make sense. One such example is from the analysis of *Tom Sawyer, Detective*, which used words that refer to a homicide such as *kill* (18 times), *murder* (13 times), *killed* (13 times). Muder is a characteristic phenomenon in stories that have a detective, which means that it is reasonable to assume that it should be unique compared to all the other works in this regard. It should also be noted that the aforementioned book was the only one to fit the narrative shape of a "man in a hole" story, which increases in negative sentiment until an antagonist or opposing force is dealt with and then rapidly increases in positive sentiment.

Another problematic aspect of sentiment analysis are words that can act as intensifiers, which may not have a positive sentiment value, but rather emphasise the meaning of a word that follows it. One example of this is from the book *Tom Sawyer Abroad*, which has the word *pretty* as its most frequently used positive sentiment word. The word *pretty* was used as an intensifier in constructions such as *pretty faded, pretty soon*, *pretty cheap*, *pretty dull*, *pretty low*, *pretty good*, *pretty fast* (Twain 2009). In the

case of constructions such as *pretty good* the sentiment analysis of this thesis counts it as two positive sentiment words, which yields the right result. However, if the construction is *pretty dull*, then the first word is counted as positive and the latter as negative, which produces a suboptimal result.

Using sentiment analysis for finding general shapes of stories is a rather new field of study but can yield interesting insights for researchers. Treating textual works as data may provide researchers new insights into the work of an author. One could analyse the life work of one author and portray the differences in the works throughout his or her life. Additionally, the aforementioned method could be used to compare the works of authors from similar genres to see if the general shapes of stories are more similar or more different than those of authors from different genres.

One interesting result of the present thesis is the fact that the "Cinderella" story curve was used in three out of the five books. According to the Encyclopedia Britannica (2020) The aforementioned three books and their first publishing dates are: *The Adventures of Tom Sawyer* 1876, *Adventures of Huckleberry Finn* 1885, *The Prince and the Pauper* 1881. Out of the five books used for this analysis aforementioned three are also the earliest works. This might indicate that Twain preferred writing narratives that have the "Cinderella" shape of stories earlier in his literary career. Analysing the works of different authors from the same time period might yield interesting insights regarding what kinds of shapes of stories were commonly used for a time period. One possible way to expand on the work done for this thesis is to cross reference the sentiment words that were commonly used during the time period of Twain's writing with the words that were indexed for this thesis. This may provide new insights for corpus linguistics regarding what kinds of words were more or less frequent in literature during a given time period.

The analysis of the shape of the stories of the five books by Twain has yielded an interesting insight. All the books ended with an increase in positive sentiment word usage. This might be an indicator that Twain preferred to end a story on a positive note. Considering the fact that the books that were chosen are young-adult fiction, then it would only make sense that the endings should have more positive words in them. It might be advantageous for future researchers to analyse other genres and compare their findings in order to evaluate whether there are patterns that are specific for certain genres.

There are some shortcomings of sentiment analysis that researchers should be aware of. Sentiment analysis is not ideal for finding all possible expressions of sentiment within a text because of characteristics such as slang words, irony, dialect specific words, sentence structure. It should be mentioned that manual analysis is also imperfect and it might be more difficult to detect errors made by a human than errors made by a program. If the shortcomings of a program are identified then one can make adjustments to the software to eliminate errors. In order to resolve issues regarding the aforementioned aspects of a text, a much more comprehensive script or algorithm would have to be created in order to properly identify the sentiment that is being expressed. Another problematic aspect is the selection of software used for analysing large data sets. During the sentiment analysis of five books by Mark Twain, it became apparent that Excel was not ideal for sorting through all the data tables and at times crashed. In addition to program crashes, data distortion was also a worryingly frequent issue. The aforementioned issues were avoided by the use of RStudio. With all that being said, the endeavour to test and research sentiment analysis is a worthwhile undertaking, because it has the potential of producing new insights and may become a new useful tool in the arsenal of literature analysis.

CONCLUSION

Treating works of fiction as textual data has the potential for providing new insights into what kind of words are frequently used within a text. In the case of the selected works by Mark Twain the most frequently used sentiment word was *pretty*, which might seem like a trivial piece of information, but one should consider how long it would have taken to find that out with manual reading. The script created for this thesis scanned through five books and created data tables and figures that provide the possibility of comparing the books to each other without having to manually go through every page. The theoretical section of this thesis provides an overview of the application of sentiment analysis in the field of humanities. In addition, this thesis provides descriptions regarding the methodology of the sentiment analysis. The aim of the analysis was to use statistical software in order to find sentiment patterns in works of literature. The sentiment analysis was done with RStudio and packages that were available through it. In addition, the benefits and consequences of choosing a sentiment lexicon were discussed in order to provide information for future researchers.

Sentiment analysis is imperfect. It is important to both eliminate shortcomings in the analysis process and accept that this technology is rather new and thus it will not produce perfectly accurate results. However, if the shortcomings are clearly reported then it is still possible to comprehend what parts of the analysis are relevant.

The analysis of the five books indicated that Mark Twain might have preferred writing books of young-adult fiction that had the “Cinderella” shape of stories. In addition, there were frequent clusters of positive sentiment words near the end of all of the books that were analysed. The analysis of sentiment word frequency indicated that there are

patterns of sentiment words regarding the general theme of the story. In *The Prince and the Pauper* words such as *majesty*, *poor*, and *master* were frequently used. A similar pattern emerges from the book *Tom Sawyer Detective*, which has among its frequently used words such as *kill*, *murder*, *killed*. Analysing the percentage of sentiment words per book provided an insight into how many positive and negative sentiment words were written into the stories. The aforementioned might indicate the amount of negative emotions depicted in a story.

Sentiment analysis can yield new insights into works of fiction, because it provides a bird's-eye view of one or several works of fiction. The scope of insights depends on the research aims. In the case of this thesis, the aims were: finding the most frequent sentiment words, comparing sentiment fluctuations from beginning to end, and estimating the general positive and negative sentiment of the works that were selected. The thesis provides a data centered perspective regarding the emotional disposition of the selected works.

LIST OF REFERENCES

Brand, O. Charlotte, Alberto Acerbi, Alex Mesoudi. 2019. Cultural evolution of emotional expression in 50 years of song lyrics. Available at https://www.cambridge.org/core/journals/evolutionary-human-sciences/article/cultural-evolution-of-emotional-expression-in-50-years-of-song-lyrics/E6E64C02BDB0480DB13B8B6BB7DFF598/core-reader, accessed on March 21, 2020.

Cambridge Dictionary. 2020. API. Available at https://dictionary.cambridge.org/dictionary/english/api, accessed on March 21, 2020.

Encyclopedia Britannica. 2020. Mark Twain. Available at https://www.britannica.com/biography/Mark-Twain, accessed on May 12, 2020.

Flood, Aliso. 2016. Three, six or 36: how many basic plots are there in all stories ever written? Available at https://www.theguardian.com/books/booksblog/2016/jul/13/three-six-or-36-how-many-basic-plots-are-there-in-all-stories-ever-written, accessed on April 23, 2020.

Heuser, Ryan, Franco Moretti, Erik Steiner. 2016. The Emotions of London. Available at https://litlab.stanford.edu/LiteraryLabPamphlet13.pdf, accessed on March 21, 2020.

Jockers, L. Matthew. 2015. Revealing Sentiment and Plot Arcs with the Syuzhet Package. Available at http://www.matthewjockers.net/2015/02/02/syuzhet/, accessed on March 21, 2020.

Johansson, Emil. 2012. Lord of the Rings Project. Available at http://lotrproject.com/statistics/books/sentimentanalysis, accessed on March 21, 2020.

Lawlor, Anna. 2014. Five inconvenient truths about social data. Available at https://www.theguardian.com/media/2014/aug/04/five-inconvenient-truths-social-data-marketers, accessed on May 09, 2020.

Liu, Bing. 2015. Sentiment Analysis. New York: Cambridge University Press.

Hu, Minqing, Bing Liu. 2004. Mining and summarizing customer reviews. Available at https://dl.acm.org/doi/10.1145/1014052.1014073, accessed on April 12, 2020.

Morin, Olivier, Alberto Acerbi. 2017. Birth of the cool: a two-centuries decline in emotional expression in Anglophone fiction. Available at https://pure.tue.nl/ws/files/88185435/Birth_of_the_cool_a_two_centuries_decline_in_emotional_expression_in_Anglophone_fiction.pdf, accessed on March 21, 2020.

Parksepp, Lotte. 2019. Sentiment mining in Goodreads Reviews of Classic American Novels. Available at https://dspace.ut.ee/bitstream/handle/10062/64130/Parksepp_Lotte_BA_Thesis.pdf?sequence=1&isAllowed=y, accessed April 3, 2020.

Reagan, J Andrew, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth and Peter S. Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. Available at https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0093-1, accessed March 20, 2020.

RStudio. 2020. About RStudio. Available at https://rstudio.com/about/, accessed on April 19, 2020.

SentiStrength. n.d. About SentiStrength. Available at http://sentistrength.wlv.ac.uk/#About, accessed April 03, 2020.

Silge, Julia. Robinson, David. 2019. Text Mining with R. Available at https://www.tidytextmining.com/tidytext.html, accessed on March 1, 2020.

The R Foundation. 2020. The R Foundation. Available at https://www.r-project.org/foundation/, accessed on April 19, 2020.

Twain, Mark. 2006a. The Prince and the Pauper. Available at https://www.gutenberg.org/files/1837/1837-h/1837-h.htm, accessed on April 23, 2020

Twain, Mark. 2006b. The Adventures of Tom Sawyer Available at https://www.gutenberg.org/files/74/74-h/74-h.htm, accessed on May 05, 2020

Twain, Mark. 2006c. Adventures of Huckleberry Finn. Available at https://www.gutenberg.org/files/76/76-h/76-h.htm, accessed on May 5, 2020.

Twain, Mark. 2009. Tom Sawyer Abroad. Available at https://www.gutenberg.org/files/91/91-h/91-h.htm, accessed on April 23, 2020.

Wickham, Hadley. 2014. Tidy Data. Available at https://www.jstatsoft.org/article/view/v059i10, accessed on March 1, 2020.

Appendix 1A. The script for the RStudio environment

Script for the sentiment analysis of Mark Twain. Available at https://github.com/Raevukutsu/SentimentAnalysis/blob/master/Beta%20Script%20(final), accessed on April 12, 2020.

Appendix 1B. The data tables of the most used sentiment words

GitHub folder of the tables regarding most used sentiment words. Available at https://github.com/Raevukutsu/SentimentAnalysis/tree/master/Top%20words%20data/Tables, accessed on April 13, 2020.

Appendix 1C. Data tables of curves of sentiment

GitHub folder of the tables regarding sentiment curves. Available at https://github.com/Raevukutsu/SentimentAnalysis/tree/master/Curve%20Data, accessed on April 19, 2020.

RESÜMEE

TARTU ÜLIKOOL
ANGLISTIKA OSAKOND

Roland Rae

Sentiment Analysis of Selected Works by Mark Twain with the Statistical Software R

Meelsusanalüüs valitud Mark Twain'i teostest statistikatarkvaraga R

bakalaureusetöö
2020
Lehekülgede arv: 39
Annotatsioon:

Käesolev bakalaureusetöö uurib meelsus sõnade kasutust valitud Mark Twain'i teostes, milleks olid: *The Adventures of Huckleberry Finn, The Adventures of Tom Sawyer, The Prince and the Pauper, Tom Sawyer, Abroad, Tom Sawyer, Detective*. Analüüsi teostamiseks loodi 358 realine skript programmeerimise keeles R.

Töö teoreetiline osa annab ülevaate sentimendi kasutusest humanitaaria valdkonnas. Lisaks on töös kirjeldatud analüüsi sooritamise metoodika. Sentimendi analüüs sooritati RStudio keskkonnas ning rakendati lisa tarkvarapakette, mis eelnevalt mainitud keskkonnas saadaval olid.

Töö empiiriline osa analüüsib kogutud andmeid. Valitud teoste sentimendi analüüsi sooritati kolmel viisil. Esiteks, tuvastati kõige enim kasutatud sentimendi sõnad valitud teostes. Teiseks, kasutati kogutud andmeid, et kuvada sentimendi sõnade kasutust teoste algusest lõpuni. Kolmandaks, kujutati negatiivsete ja positiivsete sentimendi väärtustega sõnade esinemise protsenti.

Märksõnad
sentiment, Mark Twain, Ameerika kirjandus, emotsioon, digihumanitaaria, RStudio,

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Roland Rae,

1. annan Tartu Ülikoolile tasuta loa (light litsentsi) enda loodud teose
Sentiment Analysis of Selected Works by Mark Twain with the Statistical Software R,
mille juhendajad on Jane Klavan ja Artjoms Šeļa.

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace'is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja Lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Roland Rae
Tartus, 26.05.2020

Autorsuse kinnitus

Kinnitan, et olen koostanud käesoleva bakalaureusetöö ise ning toonud korrektselt välja teiste autorite panuse. Töö on koostatud lähtudes Tartu Ülikooli maailma keelte ja kultuuride kolledži anglistika osakonna bakalaureusetöö nõuetest ning on kooskõlas heade akadeemiliste tavadega.

Roland Rae
Tartu, 26.05.2020