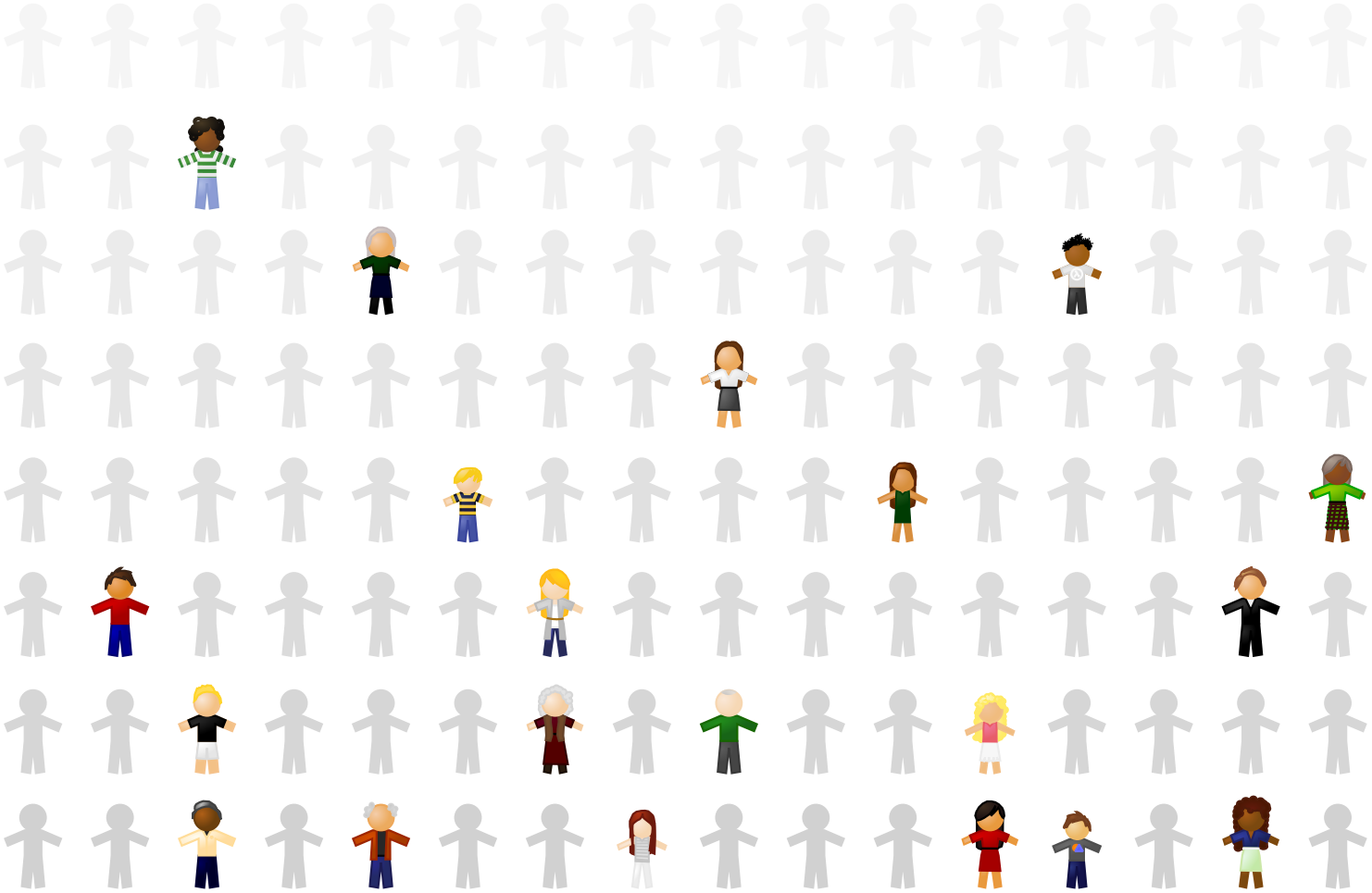


Author context extraction for interpreting the external validity of opinion mining results

July 19, 2013, O. Bloemen

Nurtured by the seemingly ever-growing amount of user-generated content publicly available on the Internet, opinion mining is a growing field of interest. Applications show promising results in brand monitoring, where the public sentiment towards product features is observed, the monitoring of sentiment towards current issues in politics, or the prediction of voting outcomes.

Despite the fact that the existence of opinion mining can be traced back to span at least a decade, the validity can be questioned. The majority of publications in this research area provide new techniques for text-analysis and/or incremental innovations for sentiment classifiers. By using labeled corpora, these innovations are typically compared to a baseline method to indicate their superior accuracy in extracting features and determining the corresponding sentiment.

For generalizability of the results, however, it is important to possess information about the context of the sample; i.e. information about the authors of the user-generated content. When this information is missing, the relation between the sample and the target population is unknown and conclusions cannot be drawn, possibly rendering opinion mining reports useless in practice.

The purpose of this thesis is to investigate whether this problem related to the external validity of opinion mining, or more specific, the generalizability of opinion mining results with respect to the review authors, indeed plays a role. Three types of threats to the external validity are identified from a literature review and examined: (1) a mismatch in demographic characteristics of the sample, (2) manipulation of online reviews, and (3) bias due to irrelevant experiences. Different methods are proposed and tested to analyze the influence of these biases on the sentiment report. Theoretical sampling confirmed existence of both a demographic and an experience bias.

Foreword

To me, this research is a great ending to my studies at the University of Twente. It allowed me to use different skills I have acquired over the years from studies and hobbies, as well as gaining new knowledge in the interesting field of data science. Starting Business Administration, I never thought I would do a master thesis as fun as this.

Throughout this research I have greatly enjoyed the discussions with supervisors Fons and Natalie. I am grateful for this research opportunity and want to thank Fons and Natalie for their enthusiasm and valuable feedback throughout the research. Next, I want to thank my girlfriend Mirte for being so patient with me, and together with my supportive friends and family, providing the sometimes necessary distraction from working on the research. I would also like to express my gratitude towards Goodreads for allowing me to use part of their data for this research.

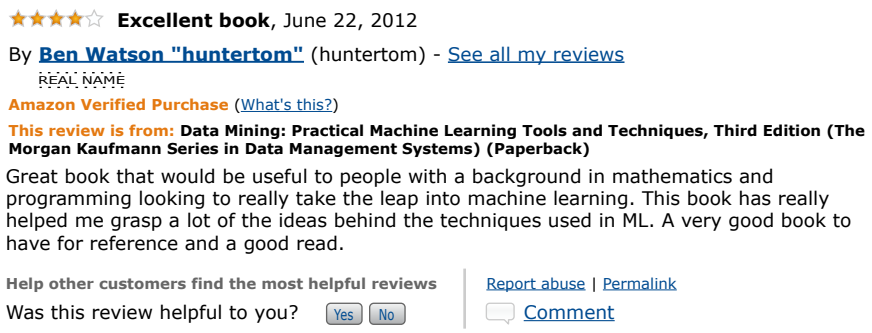
Contents

1. Introduction

Gaining knowledge of the experiences clients have with a company's and their competitors' products and/or services can be crucial for the survival of the company, as responding correctly to this information can lead to a competitive advantage (Narver and Slater, 1990). Nowadays, such brand monitoring can be assisted by using the ever growing amount of user-generated content on the Internet (Ziegler and Skubacz, 2006). Large amounts of opinions, containing potentially valuable market information, are publicly available in, for example, (micro-)blog postings, product reviews, and forums (Pang and Lee, 2008). However, due to the vast amount of available data to tap into, automation is desirable (Witten et al., 2011).

Solutions for the automated extraction of opinions come from a subsidiary of machine learning, named opinion mining. The term opinion mining was coined first by Dave et al. (2003), but the fundamental ideas for mining sentiment emerged in 2001 with the paper of Pang and Lee (2008). Depending on the broadness of the chosen scope, the roots can even be traced back several decades ago, to the classification of documents (Pang, Lee, and Vaithyanathan, 2002), as determining the sentiment of a piece of text is in essence a classification problem with classes positive, negative (Pang and Lee, 2008), and neutral (Koppel and Schler, 2006).

A sentiment classifier typically determines the polarity of the expressed opinion by comparing words in the text with words in a lexicon of which the polarity is known. Given an opinionated piece of text as in the book review of figure 1.1, such words could be "great", "helped", and "good"; indicating a positive review. Indeed, analyzing the


★★★★☆ **Excellent book**, June 22, 2012

By **Ben Watson "huntertom"** (huntertom) - See all my reviews

REAL NAME

Amazon Verified Purchase (What's this?)

This review is from: Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems) (Paperback)

Great book that would be useful to people with a background in mathematics and programming looking to really take the leap into machine learning. This book has really helped me grasp a lot of the ideas behind the techniques used in ML. A very good book to have for reference and a good read.

Help other customers find the most helpful reviews

Was this review helpful to you? Yes No

Report abuse | Permalink

Comment

Figure 1.1.: Example of an opinionated piece of user-generated content. Review over Witten et al. (2011), taken from Amazon.com, accessed on August 16, 2012.

text with opinion mining tool PATTERN[1] shows that the review is positive (0.19 on a scale from -1 to 1, indicating negative and positive sentiment respectively).

The various applications and research of opinion mining span a large domain beyond the example of books; e.g. movies (Pang, Lee, and Vaithyanathan, 2002), commercial products and services (Turney, 2002; Dave et al., 2003; Sokolova and Lapalme, 2011), determining the public sentiment of product features (Ding et al., 2008; Brun, 2011; Xu et al., 2011), and in politics the sentiment towards a political party or topic (Pang and Lee, 2008).

In the majority of opinion mining research, the dominant topic is the classification algorithm. These algorithms are continuously improved to squeeze out the last percentage increase in accuracy (Missen et al., 2010). The interpretation and validity of the results obtained from these algorithms is, however, discussed much less in the academic literature (Gayo-Avello, 2012).

If the goal of opinion mining is harvesting market or public information for decision making, it is of importance to know how the sample corresponds with the target group for which sentiment conclusions are drawn. This problem is well known in the field of psychological and sociological research methodology by external validity, which is defined by Shadish et al. (2002, p. 83) as:

> "*validity [that] concerns inferences about whether the cause-effect relationship holds over variations in persons, settings, treatments, and outcomes.*"

In opinion mining, the cause-effect relationship is of type product-sentiment or (political) topic-sentiment and the variations in persons, settings, treatments and outcomes refer to the variations between the target group of which the sentiment is measured compared to the available Internet sample. An illustration of this problem can be given from the book review in figure 1.1; if the book was written for an audience without a background in mathematics, the book might not be as good for the target group as suggested by the result found by opinion mining, which marked the review positive.

Opinion mining researchers have only recently started to acknowledge the problem of external validity in opinion mining reports (Mislove et al., 2011). Oberlander and Nowson (2006) argue that the personality of the author influences their appraisal of events and extracts four personality treats from blogs. Wu et al. (2010) acknowledge there is a problem related to customer groups and propose a visualization of opinion mining results including customer groups. In response to the many opinion mining publications based on Twitter, Mislove et al. (2011) researched the demographics of U.S. Twitter users and found a highly non-uniform Twitter population. Gayo-Avello (2012) argues that this skewness in demographics of the sample contributes to failures in predicting electoral outcomes and encourages research towards the validity of the data used in opinion mining.

[1] PATTERN 2.3 by Smedt and Daelemans (2012), available at CLiPS.ua.be/pages/Pattern, using their default English lexicon.

This research attempts to fill the gap in opinion mining research with respect to generalization of the results by taking the *context of the author* into account. The gap in opinion mining research makes this subject exploratory in nature; no research known to the author has provided a model that made it possible to investigate problems related to external validity of opinion mining results. Moreover, no opinion mining system was known that possesses the necessary capabilities to investigate the external validity of opinion mining results.

To create and test an opinion mining system that provides the necessary information to judge the external validity, as well as incorporating this information in opinion mining research, a product design research was conducted. Walls et al. (1992) describe that a product design research consists of four components: (1) kernel theories, (2) meta-requirements, (3) meta-design, and (4) testable design product hypotheses. A kernel theory is a theory borrowed from another research discipline and translated to the discipline of interest. From this kernel theory meta-requirements are defined which set a class of goals (or problems) that have to be satisfied by the product. The meta-design aims to describe artifacts that fulfill these meta-requirements. Finally, a test is conducted in final phase to check whether the meta-design indeed satisfies the meta-requirements (Walls et al., 1992).

The four components provide clear step-wise approach to (1) apply and test the ideas of external validity in opinion mining research and (2) create a system that provides crucial information related to the external validity of opinion mining research. The kernel theory is external validity, borrowed from social sciences. Application of this theory presents three different problems related to the author that could come into play when performing opinion mining research; providing *empirical propositions* that have to be tested. In order to be able to test these empirical propositions, indicators that describe these problems are needed together with the sentiment. These indicators are found using a structured literature review and defined as meta-requirements for the opinion mining system. Different methods to collect the indicators are found in academic literature. From these, a selection is made for the meta-design of the system, including *design propositions* to test the effectiveness of the methods. The testing of the resulting system is divided into two parts, (1) the testing of empirical propositions for the importance of external validity in opinion mining research and (2) the testing of design propositions for correct extraction of the indicators.

The structure of this thesis is analogous to the four components described for product design research. The next chapter discusses the kernel theory and its application to opinion mining research. Chapter 3 presents findings of the structured literature review to define necessary indicators in meta-requirements. Chapter 4 lists different methods to collect these indicators. A selection of these methods is made and presented together with the resulting design propositions. Chapter 5 describes different cases that are chosen by theoretical sampling to find evidence for the different biases. Chapter 6 presents the results of using these cases to test both the design and empirical propositions. The last chapter provides an overview of the conclusions of this research together with a discussion of the limitations, implications, and suggestions for future research.

2. Author context kernel theory and empirical propositions

Whether the results obtained by opinion mining can be extrapolated or generalized to the target audience depends on the representativeness of the sample. To analyze problems that could occur with the representativeness, the theory of external validity from social sciences is investigated in detail. Different forms of possible biases emerge when the external validity theory is applied to opinion mining, which are translated to empirical propositions at the end of this chapter.

Since external validity is a well known concept within social sciences, the information is sought in anthologies of performing social research; Shadish et al. (2002) and Babbie (2007). While Babbie (2007, p.233–234) only gives an example of previous work by Shadish et al. (2002), in Shadish et al. (2002, p. 86–90) five threats to external validity are presented: (T1) interaction of the causal relationship with units, (T2) interaction of the causal relationship over treatment variations, (T3) interaction of the causal relationship with outcomes, (T4) interaction of the causal relationship with settings, and (T5) context-dependent mediation.

The first threat (T1) reflects to properties of the units, for example the gender and ethnicity of people, and how they relate to the causal relationship. Furthermore, it is noticed that the motives of people for participating in a study may vary (Shadish et al., 2002). The second threat (T2) relates to differences in treatments. A found relationship might not hold in combination with other treatments or variations of the treatment. Examples could be a different version of a car or an influencing factor as rise in fuel prices (Shadish et al., 2002). The third threat (T3) implies that the findings of a specific research cannot be extrapolated to different outcomes; i.e. different definitions of outcomes cannot always be account for. Shadish et al. (2002, p. 88) gives as example the effectiveness of a medical treatment. This could be measured in quality of life, 5-year metastasis-free survival, or overall survival. These outcomes differ significantly and cannot be generalized to each other. The fourth threat (T4) states that the research setting can have an influence on the results; i.e. the setting or environment could introduce a bias. The results from testing the effectiveness of a new drug could differ between first and third world countries due to different health hazards in the environment (Shadish et al., 2002). The last threat (T5) is related to the way the causal relationship works. The path that explains the causal relationship can be different across various settings (Shadish et al., 2002).

Application of these five threats to external validity in opinion mining research reveals

possible problems with opinion mining reports. Two problems can be identified from the first type of threat (T1), leading to possible biases in the results. The first form of possible bias is due to a mismatch in demographic properties of the sample and the target audience (B1). E.g. if the researcher is interested in public opinion of a population towards a specific topic, the authors of the reviews from which the sentiment is extracted must reflect this population. The next problem lies in the motivation of creating user-generated content. Reviews can be written to purposely influence public sentiment, i.e. manipulating the perceived sentiment (B2). An example for such a motivation could be to increase sales for a specific item by posting positive reviews.

The second type of threat (T2) introduces a problem related to personal experiences of the author, i.e. the sentiment is influenced by events (B3). Examples include a review author that has a negative opinion due to certain problems with an old product that would not occur in the new version, or authors that had negative experiences due to a force majeure as a power outage.

The third threat (T3) relates to the type of information that is extracted by the opinion mining tool. Problems occur, for example, if conclusions are to be drawn about the sentiment with respect to specific features of the product and only the overall sentiment is measured. This is a type of error that lies within the opinion mining algorithm and therefore considered to be outside the scope of this research. Furthermore, the specific problem given in the example is already solved, see for example Ding et al. (2008); Brun (2011) and Xu et al. (2011).

The fourth threat (T4) describes the importance of the research setting when generalizing the findings. In opinion mining research, the setting of the website(s) from which the reviews are mined could be troublesome for generalization. For instance, mining an online forum for Apple products to determine the sentiment regarding Samsung products is expected to yield different results than when mining an Android forum. While this could provide a problem in opinion mining research, here the focus is on biases related to the author of reviews instead of possible biases due to the medium (e.g. a certain forum, Twitter, Facebook) from which the reviews are mined.

The last threat (T5) described by Shadish et al. (2002) relates to the causal path that links analysis of the review texts to the sentiment of the author. These paths are typically described by features found using a machine learning algorithm. The majority of publications in opinion mining research concerns with refinement of these used algorithms (Missen et al., 2010), one could see this as a refinement in finding the correct path. As with the third threat, this threat lies within the opinion mining algorithm that is being used. Furthermore, this topic is well investigated by researchers and therefore it is deemed outside the scope of this research.

Table 2.1 presents an overview of the previously discussed relations between threats to external validity in social sciences and possible problems in opinion mining research. As the scope of this research is on interpretation of opinion mining reports, only the possible biases due to (B1) a demographic mismatch, (B2) manipulation of reviews, and (B3) experienced events are investigated. The other possible problems found from external validity applied to opinion mining are not considered here because of our focus

on the importance of the author's context with respect to their opinion.

Table 2.1.: Overview of how threats to external validity in social sciences are translated to opinion mining research.

	Threat to external validity		**Occurrence in opinion mining**
T1	Interaction of causal relationship with units	B1	Bias due to demographic mismatch
		B2	Bias due to manipulated reviews
T2	Interaction of the causal relationship over treatment variables	B3	Bias due to events
T3	Interaction of the causal relationship with outcomes	*Relates to outcomes of traditional opinion mining*	
T4	Interaction of the causal relationship with settings	*Different sentiment amongst different websites*	
T5	Context-dependent mediations	*Causal path(s) of traditional opinion mining*	

A structured literature review finds support for the possible biases in opinion mining research. Articles, books, and conference proceedings are sought that contain information about the mining or analysis of opinions or sentiment in various search engines.

With the following query:

```
"opinion mining"
OR "sentiment analysis"
OR (
  mining
  AND (
    "social media"
    OR "user generated content"
    OR reviews OR blog* OR forum*
  )
),
```

a large set of relevant literature was found which was further refined by including keywords targeting the problem area in opinion mining with respect to external validity. The following query is used to narrow the previous result set:

```
"external validity"
OR generali*
OR sample
OR noise
OR bias.
```

Inclusion of these terms reveals the existence of a gap in research. While some papers provide evidence for the influence of one or more bias sources in opinion mining (Greenberg, 2001; Oberlander and Nowson, 2006; Stone and Richtel, 2007; Jindal and Liu, 2008;

Thelwall et al., 2009; Dijck, 2009; Ye et al., 2009; Missen et al., 2010; Wu et al., 2010; Das and Bandyopadhyay, 2011; Gayo-Avello et al., 2011; Mislove et al., 2011; Wang and Xue, 2011; Hu, Bose, Koh, et al., 2012), the importance of these factors is only mentioned noteworthy by Gayo-Avello (2012) and Mislove et al. (2011).

2.1. Demographic properties of the sample

The importance of matching demographics between the sample and target audience is seen in research on Internet surveys. Both Meyerson and Tryon (2003) and Ross et al. (2005) have compared Internet surveys to their offline counterpart, traditional surveys. Meyerson and Tryon (2003) observed skewed demographics in their Internet sample. To compare their Internet survey with the results of the traditional survey they first selected a subset of the Internet sample that matched the demographics of the sample from the traditional survey. A similar caveat is present in the study of Ross et al. (2005). While they do not select a subset, they analyze the results more qualitatively and explain differences between the studies in terms of demographic differences. Both studies find good agreement with field research, but only after correcting for the skewed demographics of the Internet sample.

Meyerson and Tryon (2003) and Ross et al. (2005) show the importance of having a representative sample before drawing conclusions. Other studies regarding social media show that the demographics of users is not parallel to the population. On MySpace, for example, females dominate over males and younger people are overrepresented (Caverlee and Webb, 2008; Thelwall, 2008; Pfeil et al., 2009). On Twitter, Mislove et al. (2011, p. 21) conclude that "Twitter users significantly overrepresent the densely population regions of the U.S., are predominantly male, and represent a highly non-uniform sample of the overall race/ethnicity distribution". Mislove et al. (2011) acknowledge the implications of their findings for research and encourage researchers to incorporate the use of demographics in Twitter-related research.

An example showing the importance of incorporating demographics is given by customer reviews on Booking.com. Here, hotel ratings are accumulated for different groups of customers. An excerpt of the scores for Hotel 65 is given in table 2.2. In the table, a deviation of 0.8 points between a "Group of friends" (average score of 7.5) and "Young couples" (average score of 6.7) is seen. This deviation could be explained in terms of different expectations and priorities; a hotel might focus solely on a specific market segment. In the case of Hotel 65, the target segment might be "Group of friends", making the ratings appear a lot better than when the target segment would be "Young couples". Suppose opinion mining is performed on these reviews while the target segment being only "Group of friends". An opinion miner not incorporating these demographics will use 301 reviews that do not belong to the target segment. As can be seen from table 2.2, this will result in a bias and creates a misleading view of the actual opinion of the target audience.

Table 2.2.: Ratings for Hotel 65 in London. The data is extracted from Booking.com on June 16th, 2012.

Audience	**Sample size**	**Score**
Families with older children	48	6.9
Families with young children	23	6.9
Mature couples	32	6.9
Group of friends	**103**	**7.5**
Solo travelers	120	7.0
Young couples	**78**	**6.7**
Total size and average score:	404	7.0

2.2. Manipulation of online reviews

Dellarocas (2006) argues that the anonymity that goes together with user-generated content, combined with the growing influence of online reviews on consumer behavior, gives stakeholders incentives to manipulate online reviews. An example of this is a book publisher that tries to increase the sales of their book. To do so, the publisher can write positive reviews for their book while at the same time writing negative reviews for competing books. A visitor of the website looking for a book can then be influenced by the manipulative reviews, thinking that the book of the manipulating publisher is the best and decide to buy that book. The article of Stone and Richtel (2007) in The New York Times even links this type of online manipulation directly to the highest managerial levels of corporations.

Identification of manipulated reviews is an active topic in research. Hu, Bose, Koh, et al. (2012) show that manipulation is a serious problem, and reveal that just above 10% of the books on Amazon.com have manipulated reviews written for them. As the manipulation takes place by both writing positive reviews for own products and negative reviews for competing products (Jindal and Liu, 2008), including the manipulated reviews can give both positive or negative biases in opinion mining results.

2.3. Reviews influenced by events

Missen et al. (2010) conclude that changes of public sentiment in time can be due to demographic profiles of the posters, but they also mention the possibility that the sentiment of the blog posts is affected by the events experienced by the poster (p. 274). Similar conclusions can be drawn from the research of Das, Bandyopadhyay, and Gambäck (2012). They apply natural language processing techniques to extract more detail about the topic discussed in the review. They identify locations where, the times at, and a description of what happened that led them to a certain opinion about the topic. The idea of explaining sentiments by events is in agreement with the context model of Greenberg (2001). He stresses the changing nature of context as participants are continuously subjected to events that influence and form their opinion.

Not all these events and experiences that form the opinion of a review author might be of interest to the opinion mining researcher. An example of this is given by a hotel that recently finished a renovation. During the renovation people can be negatively influenced by nuisance due to construction works. When they write their review, this negative experience may reflect in their review. However, if the renovation is completed, opinions influenced by construction nuisance are not relevant anymore. In traditional opinion mining research this is not taken into account, possibly resulting in a more negative sentiment score compared to exclusion of people influenced by the renovation. Likewise, sentiment may be positively influenced by e.g. the presence of a famous musician at the hotel, which will probably not be relevant for you.

2.4. Empirical propositions

The literature study suggest that previously mentioned types of possible biases indeed could prove to be a problem in opinion mining research. Four different empirical propositions are defined to test the whether these biases are seen in practice. For each biases form an empirical proposition is defined. For the event bias a second proposition is added to observe if the event attracts a different audience, in which case a link could exist between B1 and B3. The propositions will be tested in the empirical part of this research and are shown in table 2.3.

Table 2.3.: Overview of the empirical propositions and corresponding bias form.

		Bias form		
	Empirical proposition	**B1**	**B2**	**B3**
EP1.1	Demographic variables can explain differences in sentiment.	✓		
EP2.1	Manipulation can explain differences in sentiment.		✓	
EP3.1	Events can explain differences in sentiment.			✓
EP3.2	Events can attract a different population.	✓		✓

3. Meta-requirements for author context identification

Determining the effect of the different external validity biases requires the availability of characteristic indicators for each bias form. These indicators are sought in academic literature and provide a set of meta-requirements which the opinion mining system has to satisfy.

3.1. Demographic bias requirements

Demographic biases in a sample imply that the representativeness of the sample for the target population is off. The relevant demographic characteristics of the target population have to be matched by the sample (Meyerson and Tryon, 2003; Ross et al., 2005; Hamilton and Bowers, 2006).

Certain variables as "gender", "age", "education" are often used and given in examples of research methodology (Babbie, 2007). To find the relevant bias indicators, these variables are included in a structured literature search, together with possible terms for the problem due to mismatch in demographics:

```
(
  "sample bias"
  OR "control variables"
  OR generali*
)
AND gender
AND age
AND education.
```

The returned result set does not appear to describe the topic related of a demographic mismatch between the sample and target population, but rather shows application of the variables. Table 4.1 provides an overview of publications and the information they extract related to demographic indicators.

The commonly used variables in social sciences expressed in Babbie (2007) include "gender", "age", "location", and "education". These variables were also most often found in literature related to opinion mining and will be used here. Lesser used variables as "personality" or "ethnicity" can be of influence for a specific type of opinion mining research but are excluded here because of the explorative nature of the research.

The chosen set of indicators is expected to be sufficient for showing difference in sentiment amongst different groups, thereby proving the possibility of a bias due to

demographic properties of the sample. Furthermore, their common usage in literature makes these indicators a good starting point for future research to elaborate on.

In summary, the following indicators are defined to monitor the demographic properties of the sample:

Gender A nominal measure consisting of the attributes "female" and "male", indicating the sex of the review author.

Age The length of existence of the author at time of creation of the user-generated content in years. The time of creation is important as the author could have aged significantly since posting the review.

Location A characteristic physical place belonging to the author, e.g. the city, village, or town, in which the author is a resident, in terms of longitude and latitude.

Education The education level of the author at time of writing the review using an ordinal measure consisting of "no education", "elementary", "high school", and "college", named in order.

3.2. Manipulation bias requirements

Manipulated reviews give a false impression of the overall sentiment. To examine the effect of manipulation in opinion mining these false reviews have to be detected. A structured literature review is used to find indicators for manipulated reviews. The following search query was issued to academic search engines to find relevant papers about online manipulation:

```
(
  manipulat*
  OR spam
)
AND
(
 review*
 OR "user-generated content"
).
```

In a later stage terms regarding identification and opinion mining were added to narrow the result set:

```
...
AND
(
  detect*
  OR identify*
),
```

and

```
...
AND
(
  "opinion mining"
  OR "sentiment analysis"
).
```

From the literature review the papers in table 4.3 were selected. While the variables that were used to detect manipulation differ across the papers, in general the idea is to search for outliers, i.e. differences in online behavior compared to genuine users.

Literature notes that manipulated reviews have a higher chance of being near-duplicates of each other. The crafting of all unique manipulating reviews is labor intensive and therefore expensive. A relatively easier method of manipulation is by using a template text that can be filled with details and posted everywhere. By using such a template, the resulting reviews are roughly similar, leading to near-duplicates (Balaguer and Rosso, 2011). Jindal and Liu (2008) even go as far to split up their corpus in a near-duplicate "manipulated reviews" corpus and a "genuine reviews" corpus to find characteristics of manipulated reviews.

One of the findings of Jindal and Liu (2008) is that the near-duplicate reviews appear to be more often among the first being posted. From the standpoint of a manipulator this makes sense, as the first review completely determines the average sentiment until another review is posted. Posting a manipulated review as the hundredth review, however, will only by able to impact the average by a single percentage. P. Lim et al. (2010) also reports the impact to be helpful in detecting manipulation of reviews.

Furthermore, as a manipulator has as goal to influence the sentiment, the sentiment of the review posted by the manipulator is expected to be different. Most of the found literature takes a form for this metric into account, see for example Jindal and Liu (2008), P. Lim et al. (2010), and Chandy and Gu (2012). Elaborating on using the sentiment as an indicator, a manipulator might show a strong bias towards the brand for which it is manipulating (Jindal and Liu, 2008; Lu et al., 2010; P. Lim et al., 2010). Thereby even downplaying competing products (Jindal and Liu, 2008).

Following the previous discussion, the following indicators are defined to facilitate in identifying manipulative reviews:

Polarity deviation The difference in sentiment expressed by the review compared to the average sentiment for the product. A ratio where 0 indicates a sentiment score equal to the average sentiment and 1 for the maximum possible deviation.

Impact The influence the review has on the average ratings at time of posting. A ratio measure where 1 indicates that the review sets the complete average, 0.1 10% of the average, and 0.01 1% of the average.

Near-duplicate score Score that displays the maximum similarity between the given review and the other reviews in the corpus. The score is a ratio where 0 indicates all other reviews are completely different and 1 as an exact match.

Reviews per product The amount of reviews the author has posted on the same site for the same product.

Product bias In a set of competing products, the difference in average sentiment score for the product the given review is about versus the sentiment score of reviews from the author for competing products.

3.3. Event bias requirements

The opinions of people about a certain topic are formed by experiences (Greenberg, 2001). Particular experiences might be of interest, such as a new version of the product, or not of interest, such as negative sentiment due to maintenance and revision.

The following search query is used to find relevant literature to identify events from user-generated content and provide event indicators:

```
(
  event*
  OR experience*
  OR topic*
  OR trend*
)
AND
(
  extract*
  OR collect*
  OR cluster*
  OR emerg*
  OR identify*
)
AND
(
  Internet
  OR web
  OR online
  OR "user generated content"
  OR "social media"
  OR twitter
).
```

A relevant research area that identifies and clusters emerging trends is also included as the theory proposed there could possibly be transferred to event identification and clustering.

The query is further extended to specifically target opinion mining:

```
...
AND
(
  "opinion mining"
  OR "sentiment analysis"
).
```

The found literature can be divided into two groups: the first group uses natural language processing techniques to extract events from text, the second group uses similarities amongst different reviews to cluster and discover topics or trends over time. The selected papers are presented in table 4.5.

The papers concerning event extraction by using natural language processing focus on one review at the time whereby they typically try to extract the topic, who experiences the event, the sentiment effect of the experience, and a temporal aspect of the event. Examples are Inui et al., 2008 and Abe et al., 2011 who build an experience database, or Saurí and Pustejovsky (2012) who investigated the factuality of an experience (when and if the experience happened).

In the literature that applied a clustering approach incorporated, in addition to the previously mentioned properties, a measure to indicate how many reviews are affected by the event. Furthermore, clustering approaches can take advantage of meta-data next to the review texts, which allows in some cases for convenient extraction of a location for the event. See for example Becker et al. (2010) or Longueville et al. (2009), where the amount of appearances of the events in texts is used to select more accurate "larger" events. The corresponding locations are collected from context features available in their social media data source.

The following properties of an event described in this section are considered to be sufficient for this research for identification and analysis of events in relation with opinion mining:

Reach The amount of reviews that are influenced by the event.

Date and time Indication of the temporal origin of the event described by a date and time.

Location A characteristic physical place for the event in terms of longitude and latitude, likewise as for the demographic indicator.

Sentiment effect The effect the event has on the sentiment of the review authors that are influenced by the event.

Description Characteristic textual features that describe the identified event.

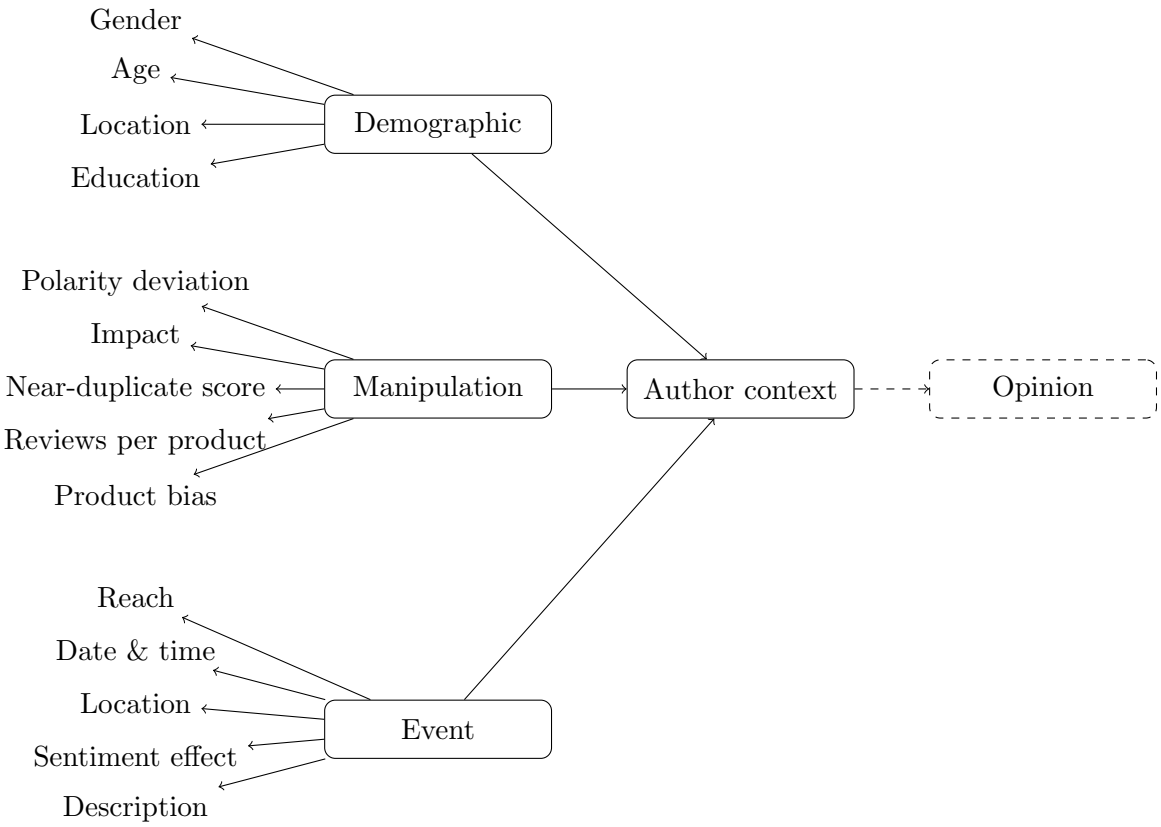


Figure 3.1.: The author context model in which an opinion is viewed. The different context contributions, demographic, event, and manipulation, relate to the possible forms of bias identified from the kernel theory.

3.4. Overview of possible biases forms and variables

Figure 3.1 shows the different aspects of how the context of the author is described in terms of the previously defined indicators. As the research is focused on the generalization of opinion mining research, the opinion derived from the opinion mining algorithm is marked by a dashed border. The task of finding the sentiment from a review text is a meta-requirement for the system, but will be delegated to opinion mining tool PATTERN[1].

For completeness, the meta-requirements are listed in table 3.1. The meta-requirements follow from the different indicators previously discussed. Different algorithms to collect these indicators are presented in the next chapter.

[1]PATTERN 2.3 by Smedt and Daelemans (2012), available at CLiPS.ua.be/pages/Pattern, using their default English lexicon.

Table 3.1.: Overview of the meta-requirements and corresponding bias form.

Bias form		**Meta-requirement**	
B1	Bias due to demographic mismatch	MR1.1	Gender
		MR1.2	Age
		MR1.3	Location
		MR1.4	Education
B2	Bias due to manipulated reviews	MR2.1	Polarity deviation
		MR2.2	Impact
		MR2.3	Near-duplicate score
		MR2.4	Reviews per product
		MR2.5	Product bias
B3	Bias due to events	MR3.1	Reach
		MR3.2	Date & time
		MR3.3	Location
		MR3.4	Sentiment effect
		MR3.5	Description

4. Meta-design for author context identification

The meta-design consists of different methods that extract the indicators defined in the meta-requirements. To find these methods first an examination is made on the (expected) available Online data or "content features". These content features are then used to guide the search for methods and the meta-design.

Typically, opinion mining collects online review texts about a certain product or topic. This text is then analyzed to derive the sentiment score of the author (Pang and Lee, 2008). But often more information is available besides the review text. The available information of a review can consist of a (1) text containing the review, (2) a score indicating the overall sentiment of the author, (3) a date and time when the review was issued, (4) a date and time when the review was last modified, (5) comments belonging to the review, (6) the username of the author, and (7) a link to the profile of the author.

In case of a link to the profile of the author, even more information might be available. Abel et al. (2010) and Balduzzi et al. (2010) examined the amount of information that is accessible on user profiles of large social network sites as Facebook, MySpace, Twitter, LinkedIn, and Flickr. Their findings indicate that profiles can give (1) an username, (2) a first/last/full name, (3) a profile photo, (4) a homepage, (5) a location, (6) an email-address, (7) their friends, (8) their age, (9) their gender, (10) their job, (11) their education, (12) their relationship status, and (13) their sexual preference. However, the available information differs between various social network sites (Abel et al., 2010; Balduzzi et al., 2010).

4.1. Extracting demographic indicators

The search for indicator extraction algorithms starts with the first bias type, bias due to a demographic mismatch. The goal is to search for the indicators gender (MR1.1), age (MR1.2), location (MR1.3), and education (MR1.4).

From observing the expected available content features, the term stylometry is incorporated as this term reflects the study area of linguistics; i.e. literature is sought that describes how review texts can be analyzed to extract the demographic indicators. The following search query was constructed:

```
stylometr*
AND
(
```

```
    profil*
    OR age
    OR gender
    OR education
).
```

The indicators are included to refine the search as the interest is only in stylometry articles which include one or more of these items.

Other relevant literature is sought using a more general query related to extraction of demographic variables from Internet with respect to profiles and user-generated content:

```
(
    profil*
    OR demograph*
)
AND
(
    internet
    OR web
    OR online
    OR "user-generated content"
)
AND
(
    extract*
    OR collect*
).
```

The result of the combined literature studies, the search for indicators in the previous chapter and the search for extraction methods in this chapter, regarding demographic biases is shown in table 4.1. The table provides an overview of the papers, which indicators they identified relating to demographics, whether they provide evidence for a bias in demographics, and the type of extraction method they used. In the next subsections, these extraction methods are explained.

4.1.1. Extracting demographic indicators from user profiles

Different indicators are possibly directly available on the profile pages of the review authors. In case the necessary demographic indicator is present at the profile of the user, this indicator can be used directly. The papers of Abel et al. (2010) and Balduzzi et al. (2010) note that on some social networks all demographic variables are available, i.e. the author's gender (MR.1.1), age (MR1.2), location (MR1.3), and education (MR1.4).

The problem with profile information is, however, that the information is often self-provided and one can question whether this information is trustworthy. An illustration of the reality of this problem is given by the work of Caverlee and Webb (2008). They

Table 4.1.: Found papers related to a demographic bias in opinion mining research. (1) indicates personality traits, (2) political affiliation, and (3) ethnicity.

	Demographic indicator						**Extraction method**		
Publication	**(MR1.1) Age**	**(MR1.2) Gender**	**(MR1.3) Location**	**(MR1.4) Education**	**Other**	**Evidence for influence**	**Stylometry**	**Profile extraction**	**Name classification**
Abel et al. (2010)	✓	✓	✓	✓				✓	
Argamon, Koppel, Fine, et al. (2003)		✓			(1)		✓		
Argamon, Koppel, Pennebaker, et al. (2009)	✓	✓	✓		(1)		✓		
Balduzzi et al. (2010)	✓	✓	✓	✓				✓	
Can and Patton (2004)	✓						✓		
Caverlee and Webb (2008)	✓	✓	✓				✓	✓	
Chandramouli and Subbalakshmi (2009)		✓					✓		
Cheng et al. (2011)		✓					✓		
Dahllof (2012)	✓	✓			(2)		✓		
Estival et al. (2008)	✓	✓	✓	✓			✓		
Filippova (2012)	✓	✓					✓		
Gayo-Avello et al. (2011)	✓	✓	✓			✓			
Gayo-Avello (2012)	✓	✓	✓			✓			
Geng et al. (2007)	✓						✓		
Goswami et al. (2009)	✓	✓					✓		
Meyerson and Tryon (2003)						✓			
Mislove et al. (2011)		✓	✓		(3)	✓		✓	✓
Missen et al. (2010)	✓	✓	✓			✓		✓	
Oberlander and Nowson (2006)					(1)	✓	✓		
Peersman et al. (2011)	✓	✓					✓		
Pfeil et al. (2009)						✓		✓	
Prasath (2010)	✓	✓					✓		
Ross et al. (2005)						✓			
Sarawgi et al. (2011)		✓					✓		
Singh and Tomar (2009)	✓	✓	✓	✓				✓	
Teo and V. K. G. Lim (2000)	✓					✓			
Thelwall (2008)	✓	✓				✓		✓	
Thelwall et al. (2009)	✓	✓				✓		✓	
Wu et al. (2010)	✓	✓	✓	✓		✓		✓	

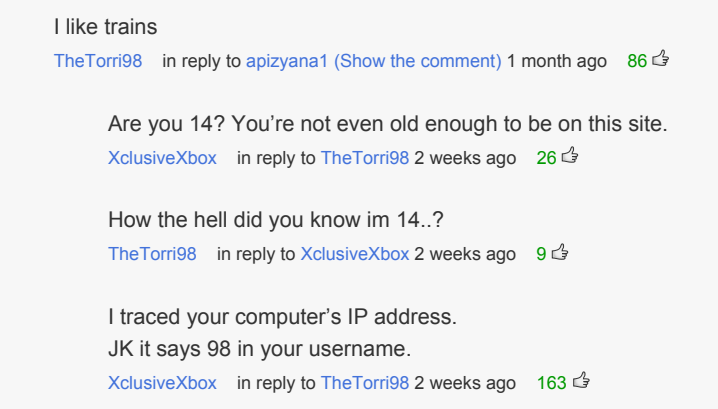


Figure 4.1.: An example of an username incorporating demographic information. Source: `YouTube.com/watch?v=2Leqeo8nIK8` (accessed July 2, 2012).

examined who used MySpace and how they used it. When they plotted an age distribution of nearly a million crawled profiles, unexpected peaks occurred around ages 69 and 100. Using text-analysis, they identified that the groups of users that provide the age 69, 99, 100, and 101, are most similar to each other in writing style, followed by people in the beginning of their 30s. One would suspect that the writing style of people is closest to people of approximately the same age. Hence, while the actual ages of the users are unknown, these findings suggest that the peaks occur due to false profile information.

4.1.2. Demographic indicators from names

"*What's in a name*?" Contrary to Julliet's notion in the tragedy written by Shakespeare, names can contain information. Usernames are deliberately chosen by individuals, and therefore are expected to hold information about this individual. Furthermore, given names and family names can reflect gender and ethnicity (Mislove et al., 2011).

An example of this is given in figure 4.1. This is an excerpt of comments on a video on YouTube. The numbers appended to the username of "TheTorri98" indicate his or her year of birth as one can deduce from the conversation. But there is even more information in the usernames of the excerpt, the inclusion of a game console in username "XclusiveXbox" can be an indication that the user plays video games on Microsoft's X-box console.

While the latter example with respect to gaming was quite domain specific, general demographic features as gender and age are expect to be extractable from (user)names. Names, numbers, and words extracted from the usernames and/or names given on profile pages, can be used for gender and/or age classification. Think for instance of the name Jacob, which is likely to relate to a male[1]. Furthermore, the numbers in an username can have a relation with the age of the author as previously discussed.

[1]See for example `http://www.socialsecurity.gov/oact/babynames/decades/names2000s.html`

4.1.3. Demographic indicators by writing style

The review text itself can possibly be used to collect demographic indicators of the author. This information is only available implicitly but Argamon, Koppel, Pennebaker, et al. (2009) successfully analyzed the style of writing to recover demographic information of an anonymous author. They extract gender (MR1.1), age (MR1.2), native language, and neuroticism with 76.1%, 77.7%, 82.3%, and 65.7% accuracy respectively.

Argamon, Koppel, Pennebaker, et al. (2009) define two different type of features that are useful for authorship profiling: content-based and style-based features. For the content-based features they express problematic implementation as these features can be influenced by the writing situation. This problem can make this feature domain dependent when the features are not chosen properly. For example, one might find that "soccer" is a feature which relates to males. If the interest is now in finding the Online sentiment of a soccer match, inclusion of "soccer" as a feature will introduce problems as the feature will be used more often by females compared to a females writing Online content in general. This possibly results in an underestimate of the amount of females in the sample. Furthermore, perhaps all the content-based features that are selected are not relevant for the research topic and they do not appear in the reviews.

To overcome these problems, one can pick only style-based features (Argamon, Koppel, Pennebaker, et al., 2009). These features only include function words, words of little semantic meaning, and typical online blog elements such as hyperlinks, images, smileys and slang.

4.1.4. More demographic indicators by coupling of accounts

Abel et al. (2010) investigated grouping of online user profiles across different websites. By using social media accounts linked to a Google profile, they find that the amount of available content features (e.g. gender and age) vary across websites and propose to interweave profiles to enrich incomplete profiles. The interweaving of profiles as performed by Abel et al. (2010) requires, however, that the profile is coupled to other online user profiles and that this information is available to the researcher.

A possible way to overcome this limitation might be found in the work of Perito et al. (2011). They showed that usernames are not particularly unique for the same person across different websites and are able to group different accounts of the same user together by matching on near-similar usernames. Application of such a technique to the opinion mining system here requires a list of most likely usernames has to be composed from a given username. From this list of usernames other profiles are sought across different websites that match these usernames, resulting in a set of candidate profiles. These candidate profiles must then be tested for similarity on the available information of the user to observe if the profile indeed concerns the same person. The information on which similarities can be found can consists of profile information (e.g. location or a profile photo) and various postings of the user. This last part is an author attribution problem and can perhaps be solved using techniques from stylometry, see for example Narayanan et al. (2012).

Furthermore, other sources inside the corporation can be accessed to gain more information and/or improve accuracy. For example inclusion of transaction history (Adomavicius and Tuzhilin, 1999), thereby coupling the Online accounts to customer relationship management systems.

4.1.5. Demographic classification from profile pictures

User profiles often go together with a profile photo. This profile photo can be used to estimate the gender, age, and ethnicity of the user. Lanitis et al. (2004), and later Geng et al. (2007), use machine learning techniques to identify characteristic features from faces that correlate to the age of people. They reported average errors in age from 3.8 to 6.8 years for Lanitis et al. (2004) and Geng et al. (2007) respectively.

Different steps have to be undertaken in such a method. First, a face has to be detected from the profile photo. Next, characteristic "landmarks" that describe a face have to be extracted to convert the face into a feature-vector which a classifier can analyze. This classifier must be trained beforehand with a collection of faces and corresponding demographic indicators gender (MR1.1) and age (MR1.2) to be able to extract these.

4.1.6. Overview of demographic indicator extraction methods

Various methods are proposed in this section to extract the demographic indicators defined in the meta-requirements. Each technique is expected to have its own strengths and weaknesses with respect to extracting indicators. A combination of the methods can be used to find all indicators, as not every method can extract all indicators. Furthermore, a level of agreement amongst different methods can provide a sense of accuracy of the demographic indicator's value.

Three methods are implemented in this thesis; the first method, profile extraction, will collect the self-provided user information. This information is used to test the empirical propositions and validate other demographic indicator extraction methods. The (user)name classification method is chosen as it is expected to be both accurate and applicable in many real life cases since names are often available. The third method, identification of gender and age by analyzing the writing style is chosen because of the promising results in literature. Furthermore, writing style analysis can be applied on existing opinion mining corpora to potentially observe if any biases exist in previous research.

The coupling of user accounts and examination of profile pictures is not implemented here because of the difficulties in implementation. Coupling of user accounts requires searching a vast amount of user profiles. Asking for permission to scrape all these profiles from different websites is considered nearly impossible within the time frame of this research. With respect to the detection of demographic indicators from profile photos, the automatic detection of faces from photos in itself is not a trivial matter, neither the extraction of characteristic features from these faces. Furthermore, the method requires a large database of photos with known demographic indicators to train the classifier.

The three chosen methods lead to different design propositions that have to be tested. An overview of these propositions is given in table 4.2.

Table 4.2.: Overview of the relation between the meta-requirements and the design propositions related to demographic bias.

		Demographic indicator			
	Design propositions	**MR1.1 Gender**	**MR1.2 Age**	**MR1.3 Location**	**MR1.4 Education**
DP1.1	Profile extraction can be used to find gender	✓			
DP1.2	Profile extraction can be used to find age		✓		
DP1.3	Profile extraction can be used to find location			✓	
DP1.4	Profile extraction can be used to find education				✓
DP1.5	(User)name analysis can be used to find gender	✓			
DP1.6	Writing style analysis can be used to find gender	✓			
DP1.7	Writing style analysis can be used to find age		✓		

4.2. Extracting manipulation indicators

Table 4.3 gives an overview given of the different papers that were found in the literature search to find indicators to identify suspicious reviews. Here it is discussed how these indicators are found from the available Online content features and sentiment score of the review text.

Table 4.3.: Overview of the found literature regarding the detection of online review manipulation.

	Manipulation indicator				
Publication	**(MR2.1) Polarity deviation**	**(MR2.2) Impact**	**(MR2.3) Near-duplicate score**	**(MR2.4) Reviews per product**	**(MR2.5) Product bias**
Balaguer and Rosso (2011)			✓		
Chandy and Gu (2012)	✓			✓	
Fawcett and Provost (1996)				✓	✓
Hu, Bose, Gao, et al. (2011)	✓				
Hu, Bose, Koh, et al. (2012)				✓	
Jindal and Liu (2008)	✓	✓	✓	✓	✓
P. Lim et al. (2010)	✓	✓		✓	✓
Lu et al. (2010)				✓	✓
Mukherjee et al. (2012)	✓			✓	✓
Stone and Richtel (2007)					✓
Xie et al. (2012)		✓		✓	

4.2.1. Calculating a review's polarity deviation

The polarity deviation (MR2.1) is a measure for the difference in sentiment score of the review under investigation and the average sentiment score of all reviews in the corpus. The indicator is defined to have a value of 0 when there is no deviation and 1 for a maximum deviation.

Such an indicator can be calculated given a sentiment score function $S(\texttt{review}_i)$, which

returns the sentiment score of review i on a scale from -1 (negative) to 1 (positive). Using such a function, the polarity deviation is given by taking the absolute value of the average sentiment score minus the sentiment score of the review under investigation, divided by two. Put in to formula form, the polarity deviation is defined as:

$$pd(\texttt{review}_x) = \frac{1}{2}\text{abs}\left[\frac{1}{N}\sum_{\texttt{reviews}} S(\texttt{review}_i) - S(\texttt{review}_x)\right], \tag{4.1}$$

where N is the amount of reviews in the corpus.

A possible problem with this definition can be the influence of the manipulated review on the average. In case of a relative large amount of manipulated reviews compared to the total amount of reviews, the average sentiment is influenced by a large factor. This potentially decreases the polarity deviation since the average sentiment score of all reviews, including the manipulated ones, is used.

4.2.2. Impact of a review

Different researchers have found that manipulative reviews have a higher chance to be posted amongst the first reviews (Jindal and Liu, 2008; P. Lim et al., 2010; Mukherjee et al., 2012). Being amongst the first allows the manipulators to have a relatively larger impact on the average rating of the product at time of posting (P. Lim et al., 2010).

While the literature did not specify a definition to measure this dimension for suspicious reviews, here impact is introduced as the influence a review has on the average sentiment at the time of posting. E.g. when the review was posted, it influences the overall sentiment score by x%. Following this description, the impact factor is defined as:

$$im(\texttt{review}_i) = \frac{1}{i}, \tag{4.2}$$

where i is the index of the review in the list of all reviews, ordered by time of posting and starting at 1.

4.2.3. Finding the near-duplicate score

The near-duplicates score compares one review with all other reviews in the corpus. For each review combination a similarity score is calculated, the maximum similarity score for each review is defined as the near-duplicate score.

To calculate the similarity score of two reviews, a method closely related to the "shingle" method can be used. The shingle method is used to find similar texts and works be splitting the text into different n-gram features, i.e. creates a set of word combinations of length n that appear in the review. For each review such a feature set is created. In the shingle method this set is reduced to N features to find a first selection of candidate near-duplicate texts (Broder, 2000). These text can then be analyzed in all features, finding how closely the texts match.

Here it is proposed to calculate the near-duplicate scores based on the n-gram text features. If the review is split up into all the features in the review, a comparison can

be made between the different feature sets of different reviews. A normalized "distance" between these feature sets will then provide a score on how closely they are related.

4.2.4. Amount of reviews per product

Some researchers noted that manipulators post multiple reviews from the same account on the same product page (for example Fawcett and Provost (1996), Jindal and Liu (2008), and P. Lim et al. (2010)), likely to achieve a higher impact on the average sentiment. This metric is calculated by counting the amount of reviews a user has posted for the same product. The score is equal for all reviews of the same user for the same product.

4.2.5. Finding the user's product bias

The product bias is defined as the difference between the sentiment score for the product for which the review is written versus the average sentiment score the author has given competing products. In terms of formula this is described by:

$$pb(\texttt{user review}_x) = \frac{1}{2}\left[\frac{1}{N}\sum_{\substack{\texttt{user's}\\ \texttt{brand}\,x\,\texttt{reviews}}} S(\texttt{user review}_i) - \frac{1}{M}\sum_{\substack{\texttt{user's other}\\ \texttt{brands reviews}}} S(\texttt{user review}_j)\right]. \tag{4.3}$$

An average is calculated of the sentiment scores the user has given for product x, the product that is being reviewed. From this average an average sentiment score of the competing is subtracted, resulting in the product bias.

4.2.6. Overview of manipulation detection methods

Table 4.4 provides an overview of the different design propositions that emerged from the manipulation indicators and corresponding methods. Each indicator in itself is not expected to be sufficient to determine whether a review is manipulated or genuine. The indicators can be used to help finding manipulative reviews by pointing out suspicious reviews. For example, a first review is more likely to be manipulated than the hundredth review (Jindal and Liu, 2008). If this same review text is also posted multiple times, this could well be a manipulated review. For this reason, DP2.6 is introduced to observe the combined effect of the indicators.

Table 4.4.: Overview of the relation between the meta-requirements and the design propositions related to manipulation bias.

		Manipulation indicator				
	Design propositions	**MR2.1 Polarity deviation**	**MR2.2 Impact**	**MR2.3 Near-duplicate score**	**MR2.4 Reviews per product**	**MR2.5 Product bias**
DP2.1	The polarity deviation helps in finding manipulation	✓				
DP2.2	The impact helps in finding manipulation		✓			
DP2.3	The near-duplicate score helps in finding manipulation			✓		
DP2.4	The reviews per product helps in finding manipulation				✓	
DP2.5	The product bias helps in finding manipulation					✓
DP2.6	The indicators combined can find manipulation	✓	✓	✓	✓	✓

4.3. Finding the event indicators

A structured literature review revealed two types of extraction methods for events; the first using natural language processing techniques and second using methods based on clustering principles. The selected papers regarding events are shown in table 4.5. In the table it is shown which event indicators are referenced, whether the paper provides evidence for the importance of events in opinion mining, and which method was used to identify events.

4.3.1. Extracting events by natural language processing

Specific words can be used to identify parts-of-speech where an event is described (K. C. Park et al., 2010; Abe et al., 2011). Looking for an event in such a fashion embodies natural language processing, which is perhaps easiest explained by an example. Suppose the following excerpt is taken from a hotel review:

> "*When* we were on our holiday to Spain the hotel *was* renovating its swimming pool . . . "

Here, "when" refers to the event of the author going on holiday and "was" refers to a renovation that took place. A lexicon of markers as "when" and "was" can be used to

Table 4.5.: Selected papers regarding variables that describe events, provide evidence for an event bias, and detection method. The "metadata" detection method observes changes in sentiment over time.

	Event indicator					**Detection method**		
Publication	**(MR3.1) Reach**	**(MR3.2) Date & Time**	**(MR3.3) Location**	**(MR3.4) Sentiment**	**Evidence**	**Clustering**	**Natural language processing**	**Metadata**
Abe et al. (2011)		✓		✓			✓	
Becker et al. (2010)		✓	✓			✓		
Das, Bandyopadhyay, and Gambäck (2012)		✓	✓	✓			✓	
Fukuhara et al. (2007)		✓		✓		✓		
Inui et al. (2008)	✓			✓			✓	
Ku et al. (2006)		✓					✓	
Landmann and Zuell (2007)	✓	✓				✓		
Longueville et al. (2009)	✓		✓			✓		
Mei and Zhai (2005)		✓			✓	✓		
Miao et al. (2009)		✓		✓		✓		
Min and J. C. Park (2012)		✓					✓	
Missen et al. (2010)		✓			✓			✓
K. C. Park et al. (2010)	✓	✓		✓			✓	
Saurí and Pustejovsky (2012)		✓					✓	
Tsolmon et al. (2012)		✓				✓		
Warren Liao (2005)		✓				✓		
Zhao et al. (2011)		✓		✓	✓	✓		

identify parts-of-speech that define an event from review texts.

However, identification of an event is only the beginning of event extraction (Inui et al., 2008; Abe et al., 2011). After finding the event in the text (MR3.5), the corresponding event indicators from the meta-requirements have to be found. Date and time (MR3.2) indications within (or close by) the sentence in which the event is mentioned can be used to find a specific date and time when the event took place. But even if such information is successfully found, problems can occur; notations of dates differ across cultures. Take for example the European day/month/year notation compared to the month/day/year notation used in the United States or a relative date as "yesterday".

Likewise difficulties arise for finding the location (MR3.3) and sentiment effect (MR3.4). Lexicons containing specific parts-of-speech to handle the various ways people describe the indicators that are defined in the meta-requirements are necessary for natural language processing. Furthermore, reach (MR3.1) is not expected to be extractable by using natural language processing alone, it needs a form of clustering to observe the amount of affected reviews.

4.3.2. Event identification by word frequencies

With respect to this research, an individual positive or negative experience might not be that interesting in the overall results. While the experience could provide the researcher with a new interesting insight, it will not provide strong evidence for the empirical propositions related to event bias. A significant amount of affected reviews, in the sense that it results in a non-negligible change in the public sentiment, is of interest. As this requires relatively many reviews, the difficulties in event extraction by natural language processing can be evaded by using clustering techniques.

In clustering, reviews are observed in groups which can be described by specific features. Think for example of grouping reviews that address the battery life of a notebook or the fuel consumption of a car. Events can raise attention to certain topics and create a new cluster or provide a relative growth to a cluster. Such a change can be noticeable and used to identify an event.

If many reviews are influenced by the same event, it is expected that descriptive words related to that event will have a larger usage frequency in these reviews compared to unaffected reviews; i.e. a sharp rise in fuel prices could lead to people emphasizing more on the fuel consumption of a car, leading to a rise in the usage frequency of "fuel consumption" and synonyms. Landmann and Zuell (2007) define four steps to identify such an event:

1. Composition of the reference-text corpus that represents general language usage.
2. Composition of the event-text corpus: text covering a period of interest.
3. a) Calculation of word frequencies and of relative frequencies for all words in both corpora.
   b) Calculation of differences between the relative word frequencies of specific words in the reference-text corpus versus event-text.

c) Selection of the n words with the largest differences between general language usage and the event texts.

4. Exploratory factor analysis based on the selected words.

These steps proposed by Landmann and Zuell (2007) provide a general description of the clustering methods used in the selected papers of table 4.5.

The previous step provides a set of descriptive textual features that describe the event, thereby fulfilling MR3.5. The remaining event indicators can be extracted in the following way: the amount of affected reviews (MR3.1) can be estimated by selecting the reviews that match the text features of the event, posting dates and times of the affected reviews can be used to identify the date and time when the event took place (MR3.2), the location (MR3.3) might be determined from the demographic location indicator of the affected users, and the sentiment effect (MR3.4) can be estimated by comparing the sentiment of the reviews influenced by the event to the sentiment of reviews not influenced by the event.

4.3.3. Monitoring large events by sentiment changes

The events that are particularly interesting for this research are those that influence the sentiment regarding the topic that is being investigated. If such an event would occur, a sudden change in the average sentiment can be noticed and the date and time of sentiment change (MR3.4) can be used as the temporal origin of the event (MR3.2). The work of Missen et al. (2010) gives an example of monitoring sentiment versus time.

After identification of such a sudden change in average sentiment (MR3.4), the event indicators defined in the meta-requirements must be extracted. Affected reviews can be selected by selecting all reviews within a time-span starting from the sudden change in average sentiment until the average sentiment has recovered. From these affected reviews the reach (MR3.1) is known. By looking up the demographic location indicators of the authors of the affected reviews an indication is given for the location of the event (MR3.3). A description for the event (MR3.5) has to be found by comparing word usage frequencies of the affected and unaffected reviews, in essence the clustering method by word usage frequencies applied in reverse.

4.3.4. Observing changes in review posting frequency

Another variable that can change due to an event is the posting frequency of reviews. It is thought possible that events attract a (different) population (EP3.2). As this newly attracted event population posts reviews while the regular population continues their normal posting behavior, a rise in overall review posting frequency is expected when such an event takes place.

Post-frequencies can be calculated in terms of posts per time span, e.g. posts per hour, day, or month. The frequency is effectively averaged over the time span given in the definition. Choosing a time span depends on (1) the amount of data that is available in the time span and (2) the necessary precision required by the researcher as the time span determines the coarseness of the temporal origin of the event (MR3.2).

Similarities with clustering by word usage frequencies can be seen. Instead of identifying events by observing changes in word usage frequencies, events are identified by observing changes in review posting frequency. The affected reviews are defined by the time-span from the start of the rise in posting frequency until the posting frequency is returned to a normal level. The event indicators can be extracted in the same manner as the previous method, where events are identified by sentiment changes.

4.3.5. Overview of event detection techniques

Two clustering methods are further examined in this research. The first being identification of large events by changes in word usage frequencies. Interesting about this method is that it looks for a descriptive set of features that relate to the event. The next selected method is observing changes in review posting frequency. Implementation of this method is simple compared to monitoring changes in word-frequency and it can give a quick overview of possible events in the reviews. Furthermore, it is argued that this method will identify events that attract different populations, a point of interest in this research (EP3.2).

Natural language processing and monitoring sentiment changes to identify events are both not further investigated here. Natural language processing is expected to be better suited for personal experiences instead of the larger events that alter the overall sentiment significantly. Furthermore, implementation of a natural language processing technique requires interpreting language, which is not a simple matter. Natural language processing is, however, expected to outperform determination of the sentiment effect since sentiment features belonging to the event can be extracted with better accuracy.

Identifying events from sentiment changes and from changes in posting frequency are quite similar. Review posting frequencies are favored here because of the possible identification of different populations. Both methods can suffer from complications as a sudden change in the underlying variable can occur due to multiple events at the same time. Due to this, it is possible that some reviews, hat do not correspond to the event, are used in determining the values of the event indicators.

The two chosen event identification methods give rise to different design propositions, shown in table 4.6. As previously noted, not all event indicators can be extracted directly but are extracted using the information returned from the methods. These two-step approaches in both methods are broken up into two design propositions for each method.

Table 4.6.: Overview of the relation between the meta-requirements and the design propositions related to event bias.

		Event indicator				
	Design propositions	**MR3.1 Reach**	**MR3.2 Date & time**	**MR3.3 Location**	**MR3.4 Sentiment effect**	**MR3.5 Description**
DP3.1	Word-frequencies can identify an event					✓
DP3.2	Word features can be used to find influenced reviews	✓	✓	✓	✓	
DP3.3	Post-frequencies can identify an event		✓			
DP3.4	Date and time of posting can be used to find influenced reviews	✓		✓	✓	✓

5. Research cases

The research consist of different stages, testing different implementations of the algorithms for detecting the author's context and testing the influence of the three possible biases. All different methods have to be tested for feasibility and positive contribution to the external validity of opinion mining, described in the design propositions. Furthermore, the research has to find the influence of the three bias factors that are found in literature on opinion mining results, the empirical propositions. To do so, three cases are selected by means of theoretical sampling. Each case is chosen to investigate one of the bias forms.

5.1. Book reviews on Goodreads.com

For all cases the online book community Goodreads is used. Goodreads is a popular online book review website launched in 2007. Their population now surpasses 14 million and they collectively taken more than 470 million actions regarding books. Actions being writing a review, rating the book, or mark the book as "to read". It is due to the large amounts of available reviews, together with available content features in user profiles that makes Goodreads interesting for testing the influence of the biases and proposed extraction methods.

The information can be collected using a combination of both their website and Application Programming Interface (API). The API is used to collect lists of links to all the reviews belonging to a book and extracting self-provided information from the profiles of the review authors. The website is used to extract the full review texts, which is not available through their API.

Using their API, however, only about 75% of all the reviews is accessible. This 75% is determined by an unknown internal algorithm which selects the "most popular" reviews. Their internal algorithm may introduce a bias in the collected dataset.

5.2. Dan Brown's The Da Vinci Code: Demographic bias

To investigate the demographic bias a case is sought containing a large amount of reviews together with rich profile information. Next to this, a mediocre average rating is sought rather than an extreme. Doing so will increase the likelihood of finding difference among demographic groups with respect to sentiment scores, and finding evidence for EP1.1.

For this case book reviews for The Da Vinci Code by Dan Brown are chosen. 23,526 reviews are available from 23,505 unique user accounts. Of these user accounts, 19,707 have indicated their gender (83.8%), and 8,741 have specified their age (37.2%).

Evidences for a demographic bias can already be found from the profile information, even with unknown genders females take up 14.527 user profiles of the total, making up 61.8% of the sample. This is already an overrepresentation of females and compared to the known males (22.2%) and unknown genders (16.0%) it is likely that females make up an even larger part of the sample than the stated 61.8%.

The large sample and many known genders and ages make it possible to train the machine learning algorithm to analyze the writing style per gender and age groups. Furthermore, the rich profile information gives a large validation set for both writing style analysis (DP1.6 and DP1.7) as well as extracting gender from given names (DP1.5).

5.3. Top books on the financial crisis: Manipulation bias

In literature manipulation is often described in terms of duplicates and/or linked to malicious user profiles. To investigate the possibility of manipulation a group of products competing with each other has to be found. A manipulator will have interest here in promoting its own product, while downplaying competing products.

Here a group of books related to the financial crisis of 2008 is chosen as the products of interest and the reviews are taken from Goodreads.com. A list of "most popular" books about the financial crisis is obtained from Amazon.com. Picking the top 20 books resulted in the book set listed in table 5.1.

Here the interest is in finding near-duplicate reviews (DP2.3) and promoting and discrediting reviews from a single review author between different books (DP2.5). Yet, determining whether a review is indeed created only with the purpose to manipulate is a subjective matter.

5.4. James Sallis' Drive: Event bias

As with manipulation, the influence of events on sentiment is less straight forward to examine compared to the influence of the demographic indicators which are available in the content features on the user profiles. A case is needed where it is known that an event happened and that this event most likely had an impact on the case (EP3.1). In addition to this, rich profile information is interesting as a combination of event and demographic bias could be found (EP3.2).

Here, the book Drive by James Sallis is chosen. From manual inspection of the reviews the book appeared not particularly popular compared to the movie based on the book. The book was published in 2006 while the movie came out in 2011. This time span leaves room for reviews to appear before the event happened, and therefor are known not to be influenced by the movie. Still, as the movie was premiered in 2011, it allows for review authors to be influenced by the movie and share their review to be incorporated in this research.

For this case 265 reviews are collected, all from unique user accounts. 223 have their gender specified (84.2%) and 107 shared their age (40.4%). Here an overrepresentation

in gender is less profound, males are the larger group with 143 user accounts (54.0%) compared to the 80 known females (30.2%).

With respect to event bias, the interest is in detection of the movie launch (DP3.1 and DP3.3), examining the amount of reviews that are affected by the movie (DP3.2 and DP3.4), possible changes in public sentiment regarding the event (EP3.1), and changes in population of the reviews due to the event (EP3.2).

Table 5.1.: Selected books to cover the financial crisis of 2008

Author(s)	**Title**	**Year**	**Score**	**Reviews**
Lewis, M.	The Big Short: Inside the Doomsday Machine	2009	4.18	2,216
Sorkin, A. R.	Too Big to Fail: The Inside Story of How Wall Street and Washington Fought to Save the Financial System from Crisis — and Themselves	2009	3.98	579
Paulson Jr., H. M.	On the Brink: Inside the Race to Stop the Collapse of the Global Financial System	2008	3.57	79
Morris, C. R.	The Two Trillion Dollar Meltdown: Easy Money, High Rollers, and the Great Credit Crash	2008	3.58	78
Zuckerman, G.	The Greatest Trade Ever: The Behind-the-Scenes Story of How John Paulson Defied Wall Street and Made Financial History	2009	3.91	75
Johnson, S. and Kwak, J.	13 Bankers: The Wall Street Takeover and the Next Financial Meltdown	2010	3.89	113
Lowenstein, R.	The End of Wall Street	2010	3.75	43
Cohan, W. D.	House of Cards: A Tale of Hubris and Wretched Excess on Wall Street	2009	3.68	142
Greenberg, A. C. and Singer, M.	The Rise and Fall of Bear Stearns	2010	3.09	8
McDonald, L. G. and Robinson, P.	A Colossal Failure of Common Sense: The Inside Story of the Collapse of Lehman Brothers	2009	3.80	93
Ward, V.	The Devil's Casino: Friendship, Betrayal, and the High Stakes Games Played Inside Lehman Brothers	2010	3.16	22
Tibman, J.	The Murder of Lehman Brothers: An Insider's Look at the Global Meltdown	2009	3.88	2
Kelly, K.	Street Fighters: The Last 72 Hours of Bear Stearns, the Toughest Firm on Wall Street	2009	3.56	32
Gasparino, C.	The Sellout: How Three Decades of Wall Street Greed and Government Mismanagement Destroyed the Global Financial System	2009	3.87	21
Wessel, D.	In FED We Trust: Ben Bernanke's War on the Great Panic	2009	3.65	39
Patterson, S.	The Quants: How a New Breed of Math Whizzes Conquered Wall Street and Nearly Destroyed It	2010	3.74	140
Krugman, P.	The Return of Depression Economics and the Crisis of 2008	1999*	3.80	167
Roubini, N. and Mihm, S.	Crisis Economics: A Crash Course in the Future of Finance	2010	3.83	59
Soros, G.	The New Paradigm for Financial Markets: The Crash of 2008 and What it Means:	2008	3.23	66
Foster, J. B. and Magdoff, F.	The Great Financial Crisis: Causes and Consequences	2008	3.93	7

6. Implementation and results

The various selected methods in the meta-design relate to different empirical and design propositions, each having an unique way to test. Each method and testing thereof will be handled separately together with the corresponding propositions.

6.1. Demographic bias from user profiles data

The profile extraction method is used to extract demographic indicators from the profiles of the review authors. Design propositions DP1.1 – DP1.4 describe the expectation of collecting the four indicators of gender, age, location, and education respectively. As there is no way of telling whether the self-provided information is correct, the design propositions are only tested on the availability of the data. In order to observe differences in demographic indicators versus sentiment score (EP1.1), the self provided information is assumed to be (largely) correct.

The dataset of the Da Vinci case is rich in meta-data, 83.8% of the genders and 37.2% of the ages from the review authors are known. From this available profile information it can already be concluded that females are overrepresented in the sample; 61.7% of the total sample indicated that they are female compared to 22.1% of the sample that indicated thay are male, leaving 16.2% unkown. If only the known genders are taken into account, females make up 73.6% of the population, leaving only 26.4% for the males[1]. This skewness in gender is even larger than the results Mislove et al. (2011) reported for Twitter. The self provided age in combination with gender is shown in figure 6.1. The age distribution of both genders is roughly similar and covers ages 17 until 77. The median age is at 25 for both genders.

Below the histogram of the amounts of posts per gender and age, a plot of the average sentiment score versus gender and age is shown. To overcome noise, the average sentiment score obtained by combining the reviews for all authors in age groups that span a decade, e.g. years 10–19. Around the extremes of the ages available in the dataset the amount of data that is available is sparse. For 10–19 there are 127 females and 56 males, for group 70–79 there are 33 females and only 17 males.

The sentiment of both genders towards the book seems to separate as the age of the reviewer increases. The separation goes on until males and females are a third of the male's sentiment score apart. The whole effect is seen from group 20–29 until 50–59. This region is best covered by the dataset, containing 2,943, 1,533, 639, and 424 samples for females and 1,113, 664, 314, and 181 samples for males respectively.

[1] This is equivalent to expecting a similar gender distribution for the unknown genders, i.e. extrapolating the findings.

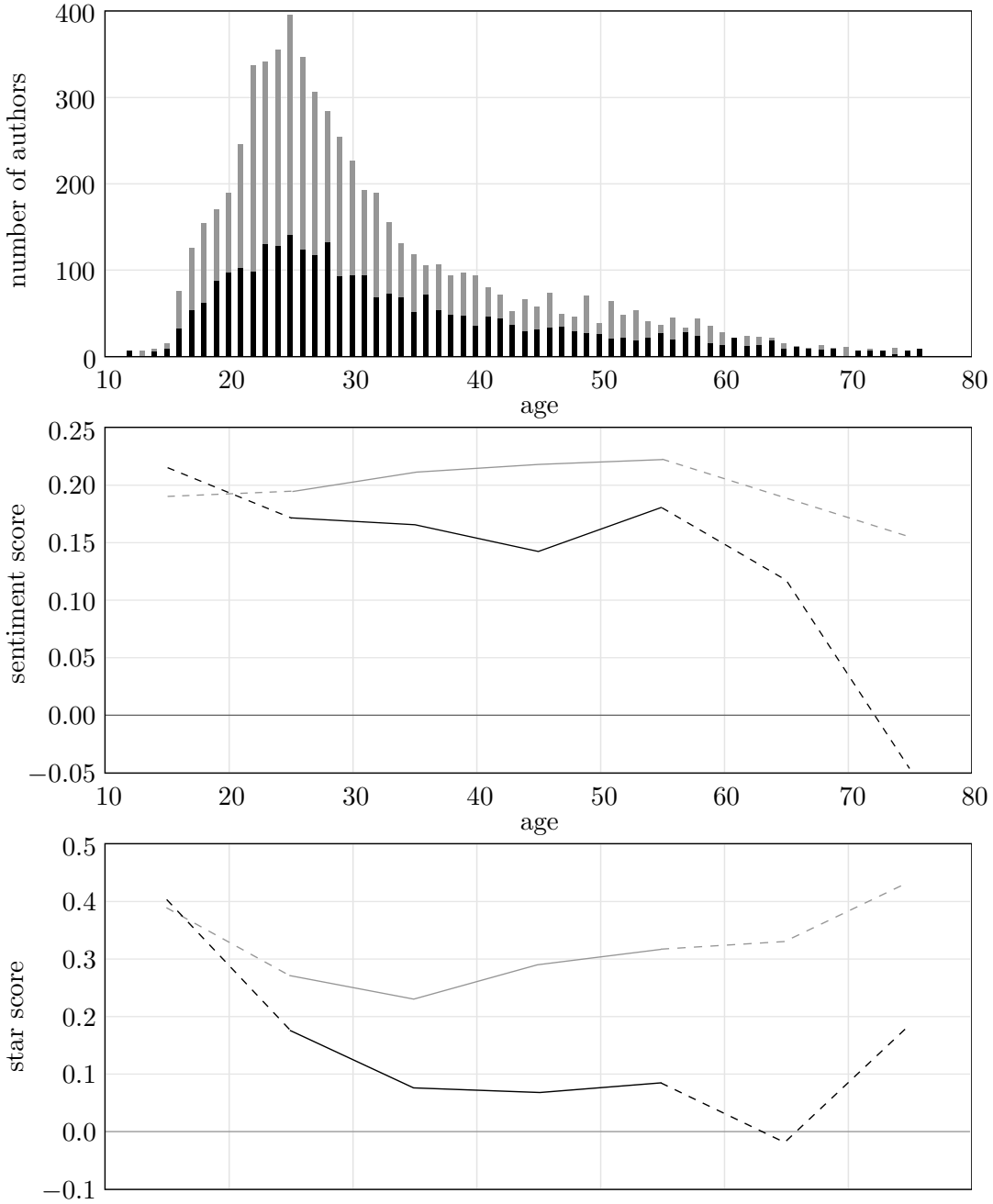


Figure 6.1.: Above the age coverage per gender from the self-provided information in the Da Vinci case, ■ and ■ being females and males respectively. Below is, first, the sentiment score based on the author's review, averaged per ten year age span; second, the normalized star score of the authors.

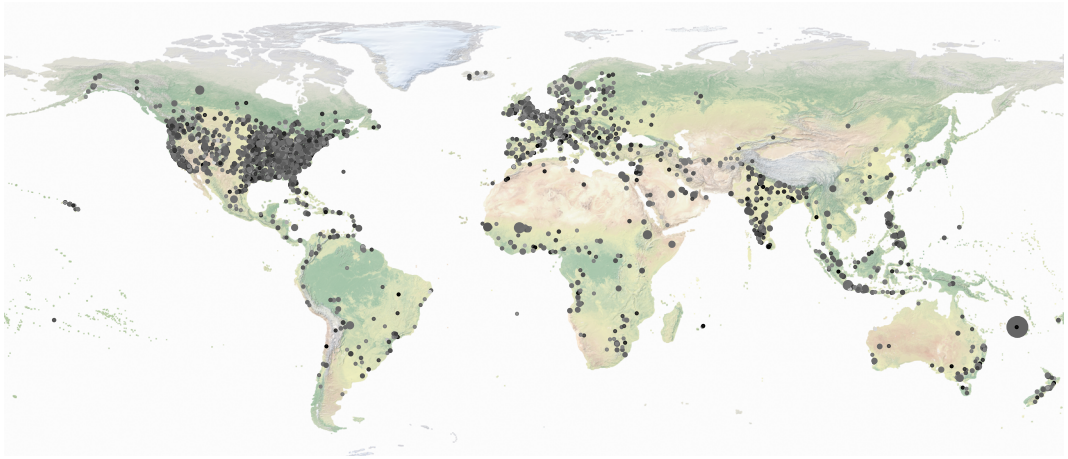


Figure 6.2.: Location and sentiment, black corresponds with negative sentiment, white with positive. The size of the circles indicate the amount of reviewers from that location. World map from Natural Earth (http://naturalearthdata.com/).

In order to check whether this sentiment distribution is an artifact of the sentiment classifier, rather than true sentiment, a plot using the "star scores" belonging to the reviews is appended below the plot of the sentiment scores. These star scores are self-provided by the review authors and span a range from 1 to 5 stars. The star score is mapped to the interval of sentiment scores ($[-1, 1]$) for comparison purposes. One can see similar characteristics in the star score graph as in the graph created by from sentiment scores. Here too, the increased separation of sentiment between genders as the subsample is older is seen, and peak around 40–49. Overall, the star score shows more extreme differences compared to the sentiment score. Drawing conclusions for absolute values is, however, dubious as there is no objective numerical basis for the amount of stars and opinion.

Furthermore, 20,664 (87.8%) of the users described a location on their profile. At time of writing there was no scientific resource at hand to translate this self-provided location to coordinates known to me. Still, to have an idea of the location of reviewers, the database of GeoNames[2] is used. The results are presented in figure 6.2. In case the self-provided location has multiple matches in the database, the location of the largest city is chosen.

The design propositions related to profile extraction questioned whether the demographic indicators could be extracted from the profile. The profile data presented here show that gender, age, and location can be found successfully on user profiles. Furthermore, the found gender and age indicators reveal different behavior in sentiment, which

[2]http://www.geonames.org/ a user-contributed collection of geographical information.

the opinion mining algorithm was able to capture the characteristics of. The education indicator was, however, not available on the profiles of Goodreads.com. These results combined lead to verification of 4 propositions, listed in table 6.1.

Table 6.1.: Overview of the propositions and results related to profile extraction.

	Design proposition	**Result**
DP1.1	Profile extraction can be used to find gender	+
DP1.2	Profile extraction can be used to find age	+
DP1.3	Profile extraction can be used to find location	+
DP1.4	Profile extraction can be used to find education	−

	Empirical proposition	**Result**
EP1.1	Demographic variables can explain differences in sentiment	+

6.2. Name analysis to find gender

One proposed method to recover the gender information for the unknown 16.0% is by analyzing the self-provided name on the profiles of the authors (section 4.1.2). The problem is a classification problem; given a name, the classifier has to decide whether it belongs to female or male category. Hence, the interest is in the accuracy and feasibility of this method for gender extraction (DP1.5). Furthermore, if the mode proves to be accurate and feasible, it can be used to analyze the importance of demographic indicators in opinion mining results (EP1.1).

Implementation of name analysis is done by searching for known name and gender relations in a given text string. For this, a database consisting of names and the odds of each name belonging to either females or males have to be known. These odds are calculated from the amount of babies born in a year with the same name and gender, a database from the Social Security Administration of the U.S. that is available to the public (Social Security Administration, 2012). Problems occur when names are given to both females and males, e.g. “Eliah” and “Kim”. There are different approaches to deal with this problem, discussed in appendix A. These approaches result in three sets of names from which the gender is known, that differ in expected accuracy; a “strict” set where every name and gender is at least 95% accurate, a “loose” set where the overall name and gender set is expected to be 95% accurate, and a “neglect” set where low name and gender occurrences are neglected before selecting a 95% accurate name and gender set.

6.2.1. Results of gender extraction by name analysis

The Da Vinci Code case is used to validate the proposed name analysis method for gender extraction. The results from application of the method using the first name in combination with the three different name-gender datasets are shown in table 6.2. Here

Table 6.2.: Results of gender prediction by analyzing the self-provided first name of the review author compared with the self provided gender. Together with the classifications of the genders, the observations and expectations (exp.), and precision (pre.) and recall (rec.) are shown.

(a) Strict 95% precise set

	female	**male**	**unknown**		**observation**			
female	9,102	80	5,337	**exp.**	12,453	91	**pre.**	0.993
male	11	3,351	1,858		7,195	0	**rec.**	0.634

(b) Loose 95% precise set

	female	**male**	**unknown**		**observation**			
female	9,720	81	4,718	**exp.**	13,238	93	**pre.**	0.993
male	12	3,518	1,690		6,408	0	**rec.**	0.674

(c) 95% precise and neglecting unknowns

	female	**male**	**unknown**		**observation**			
female	11,539	147	2,833	**exp.**	15,735	185	**pre.**	0.988
male	38	4,196	986		3,819	0	**rec.**	0.805

the gender prediction by name is compared to the self-reported gender available on from the profile of the review author. When the classifier is unable to determine the gender the case is marked as "unknown", i.e. when the name is not in the dataset.

The precision and recall of the name analysis using the three name and gender sets is calculated by comparing the correct and wrong observations of the method with the expected results. Since all review author have a gender, there cannot be a correct absence of a result and all true negatives are zero. From the observation and expectation tables the precision and recall are given by

$$\text{Pre} = \frac{tp}{tp + fp}, \text{ and} \tag{6.1}$$

$$\text{Rec} = \frac{tp}{tp + fn}, \tag{6.2}$$

where tp are the correct results (top left cell), fp the incorrect results (top right cell), and fn the missing results (bottom left cell).

Another important aspect is a possible bias towards a specific gender. People appear to be more creative with female names, resulting in an underrepresentation of females in the name and gender datasets. Table 6.3 shows the precision and recall for both genders using the three name and gender sets. The precision and recall for both genders are relatively similar across the different name and gender sets, making a bias towards a gender minimal.

Table 6.3.: Precision and recall values per gender and dataset.

	strict		**loose**		**neglect**	
	Pre	**Rec**	**Pre**	**Rec**	**Pre**	**Rec**
female	0.991	0.630	0.992	0.673	0.987	0.803
male	0.997	0.643	0.997	0.675	0.991	0.810

In figure 6.3 the sentiment score for the sample with unknown gender is shown. The sample with unknown gender shows similar characteristics as the sample from which the gender is known (dashed lines). Also visible in this figure is the difference in scores between the sentiment extract from the sentiment classifier and the self-provided star score.

One can see that the sentiment classifier has a quantized output, which is expected when one considers the inner workings of such a classifier. The peak in sentiment from the sentiment classifier around 0 is likely due to the classifier being unable to determine the sentiment, i.e. no match in the sentiment features. This can be due to short reviews, reviews that are summaries not containing an opinion, or non-English reviews.[3]

The figure shows that the distribution of sentiment scores is roughly similar to those of known gender. In fact, the averages of the sentiment scores are almost equal for the different genders. Using the neglect set, the sentiment scores are 0.21 for known females compared to 0.23 by the females discovered by name analysis. Likewise, the sentiment scores for known males are 0.17 compared to 0.18. However, a recall at 34.4% is much lower than one would expect from the validation (80.5%). Explanations for this can be people not providing a name or the fact that the name and gender set was derived from the U.S. population and not a global population, i.e. names popular outside but not within the U.S. are not taken into account correctly.

Table 6.4.: Overview of the propositions and results related to name analysis.

	Design proposition	**Result**
DP1.5	(User)name analysis can be used to find gender	+

	Empirical proposition	**Result**
EP1.1	Demographic variables can explain differences in sentiment	+

6.3. Writing style analysis for gender and age indicators

Likewise as with name analysis, analyzing writing style is tried to enhance the availability of the demographic indicators. While this method is often only used to determine the gender of the author, here it is also tried to find an age group to which the review author

[3]This same effect could explain the underestimation of sentiment score from the opinion mining algorithm compared to the star score, as seen in figure 6.1.

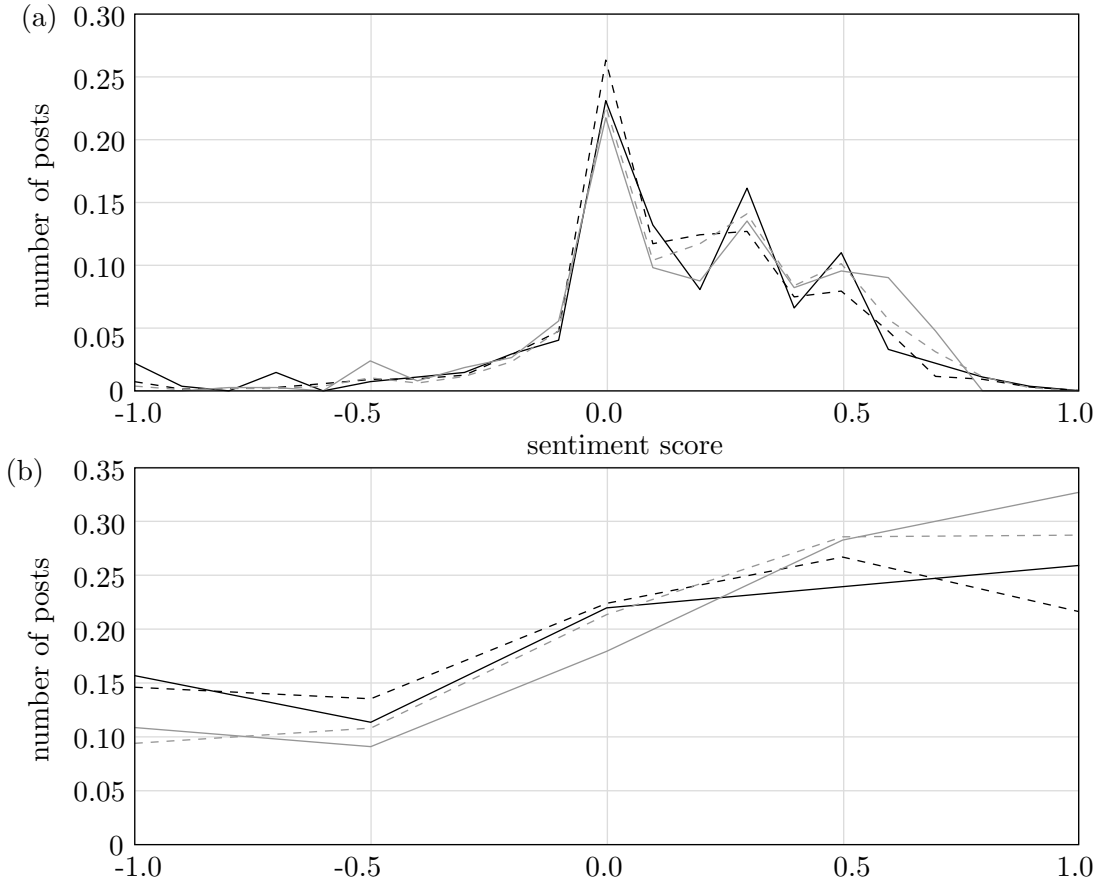


Figure 6.3.: Comparison of the sample with known gender (dashed line) and gender estimated by the name analysis method (solid line) versus sentiment by the sentiment classifier (a) and normalized self-provided star score (b). The gray lines indicate females and the black lines males.

belongs, as is also done by Argamon, Koppel, Pennebaker, et al. (2009), Peersman et al. (2011), and Filippova (2012). First the classification algorithm is discussed, followed by different training strategies to learn characteristic writing style per gender and per age group. The writing style analysis is tested against known demographic information in the Da Vinci Code case to test the associated design propositions.

6.3.1. Classification algorithm for writing style

When going through the literature, not much consensus has been reached about which classification algorithm to use for analyzing writing style. Gill and French (2007) applied writing style analysis to find personality types by Latent Semantic Analysis, Hyperspace Analogue to Language, and Pointwise Mutual Information and Information Retrieval, but all classifiers failed to perform the task at hand. Others report successes in classification of gender using Support Vector Machines (Peersman et al., 2011) and Maximum Entropy Classification (Filippova, 2012). Schler et al. (2006) uses a modified version of Winnow to outperform the baseline for age and gender classification, while in a later work the co-authors have adopted Bayesian Multinomial Regression as classification algorithm.

Due to the diverse usage of classification algorithms used in the literature, and often minimal classification improvements for more complex machine learning algorithms (Witten et al., 2011), Multinomial Naïve Bayes is used as a starting point for this research. This algorithm is capable of both binary classification for gender as well as multi-class classification for different age groups.

Multinomial Naïve Bayes examines all the features that are present in a case to calculate the most likely class the case belongs to. The probabilities of all features belonging to class X are multiplied, resulting in a probability of the case belonging to the class X:

$$P(X|E) = N!\Pi_i \frac{P(e_i|X)^{n_i}}{n_i!}, \tag{6.3}$$

where E is the evidence or the features in the specific case, N the total amount of features, e_i feature i, and n_i the amount of times feature e_i was observed in the case.

The algorithm is implemented here by picking the class with the highest probability. Choosing this implementation results in simplified equations; all factorials can be dropped as long as the probabilities for all classes and features are known.

6.3.2. Training the classifier

The algorithm has to know the probabilities of text features belonging to a class. However, there is no dataset available that links textual features, or writing style, to genders and age groups. These textual features and corresponding probabilities have to be found from a training set of texts where the gender and age are known.

The learning of features and corresponding probabilities is done here by counting the occurrences of individual words per class. Of these possible features the top 25 features

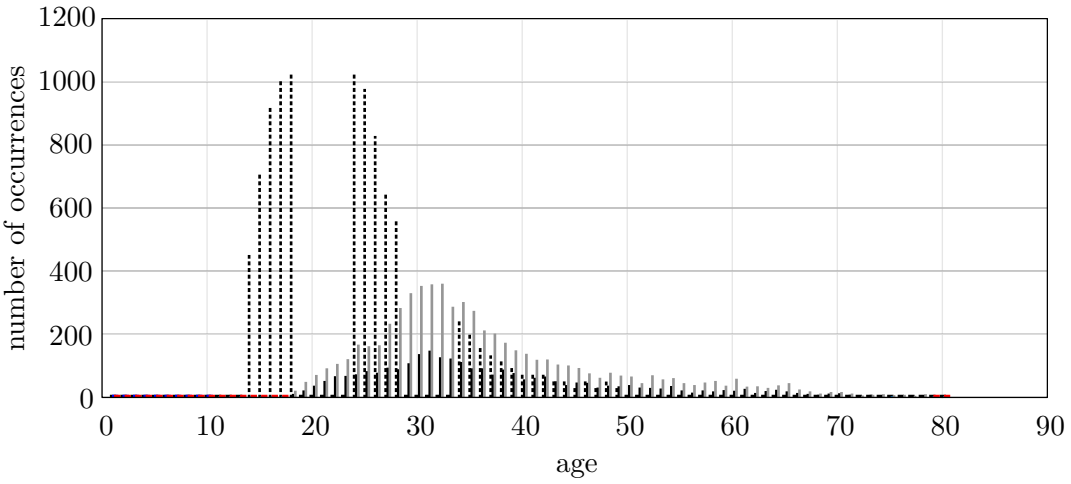


Figure 6.4.: Age and gender coverage of the blog corpus (┈┈), and the Da Vinci case corpus for females (–) and males (**–**). Note that for the blog corpus the amount of occurrences per age is equal for both genders.

with largest differences in occurrences between the classes is chosen, i.e. in a maximum entropy style. One might be tempted to select the features with largest discriminating probabilities, but doing so can result in overfitting of the feature set as features can be created that correspond to single cases. There are three different training sets used in this research.

Training from Schler's blog corpus

The blog dataset of Schler et al. (2006) is available for research and contains posts of 19,320 bloggers with self-provided gender and age. The dataset has an equal coverage for both genders, but the age distribution is not uniform, as can be seen from figure 6.4. Furthermore, the age is quantized into age groups; only reviews of authors with an age within the bins of 13–17, 23–27, and 33–47 are available.

The bins are created to remove intermediate age groups as they can introduce ambiguity since many of the blogs were written over a period of several years (Argamon, Koppel, Pennebaker, et al., 2009).[4] The corpus enforces the use the same structure in age groups. On the upside, introduction of the bins yields an age gap that can enhance application of the classification algorithm as the cases are "farther" away from each other in terms of age and therefor may posses larger differences in writing style for the classifier to exploit.

Here, the same bins will be used for the classification algorithm. For the execution this implies that reviews from authors in the missing age gap will be included in a class

[4] In this research the age indicator is defined to be at time of writing. This would overcome the problem related to age ambiguity, but the dataset from Schler et al. (2006) does not contain the necessary information to calculate this.

in which it does not belong. The classes will be named "young" for ages 13–17, "mid" for ages 23–27, and "old" for ages 33-47.

Function words from the blog corpus

Training the algorithm directly from the blog corpus of Schler et al. (2006) can cause problems due to the differences in the topics the sample blogs about versus gender and age. Content features might be extracted that negatively influence the results of analyzing the writing style to find the gender and age indicators.

To solve this problem, only function words are allowed to be features in this second training set. Argamon, Koppel, Pennebaker, et al. (2009) found that using only these features reduced the accuracy of the classifier for gender from 76.1 to 72.0% and the age classification from 77.7 to 66.9%. To find the function words, here a predefined word list is used. This list from Morales (2013) contains auxiliary verbs, conjunctions, determiners, prepositions, pronouns and quantifiers. Training the classifier now implies finding the probabilities of these function words belonging to a specific class given the blog dataset.

Training from the book reviews

The problem using a dataset that not directly corresponds to the corpus of interest is that the learned features might behave differently over the different corpora, i.e. the distribution of word usage differs between the populations. As the acquired The Da Vinci Code dataset is rich in profile data, this dataset is used as third training set. Here, both content and functional features are extracted for gender and age.

In The Da Vinci Code case both the gender and age cases are unequally distributed. This inequality is compensated for in order to find correct probabilities for feature-class combinations. The compensation takes place at candidate feature level as the length of reviews varies significantly. The compensation is implemented by scaling the candidate feature occurrences in the dataset with a factor to create an equal total feature count in all classes.

Comparison of the datasets

The top ten extracted features from the different training set are shown in table 6.5. The top features are roughly similar across the different training sets and show similarities with the features found in literature (Argamon, Koppel, Pennebaker, et al., 2009). The age bin "young" is not displayed as the Da Vinci Code case provides negligible authors within this category which makes feature extraction unreliable.

6.3.3. Results

Validation of the method is done by comparing the output of the classifier with the self-provided gender and age belonging to the reviews. The results of this validation are shown in table 6.6. Determining the precision of the classifier trained by the Da

Table 6.5.: Comparison of the top ten classification features extracted from the blog and book review dataset.

class	blog	blog functional	book review
female	I, my, me, so, am, not, and, he, him, she	I, my, me, so, he, him, she, her, like, and	I, it, was, book, this, read, loved, me, and, movie
male	the, of, a, in, is, as, this, some, for, by	the, of, a, in, as, this, for, by, on, to	of, is, the, a, that, Brown, are in, Dan, novel
mid	I, my, you, am, me, so, but, not, like, just	I, my, you, me, so, but, like, it, all, its	awesome, amazing, Dan, Brown, very, is, well, overrated, superb, novel
old	the, of, and, a, in, as, we, on, he, our	the, of, and, a, in, as, on, he, his, by	the, read, loved, this, fun, I, was, and, page, turner

Vinci Code case by using the same dataset is likely to overestimate the precision of the classifier, therefore a ten fold cross validation is used.

In ten fold cross validation the whole learning dataset is split up at random in ten different parts. Out of these ten parts, nine are used to train the classifier and the remaining one is used to test the classifier and determine the precision. This method is repeated ten times, using a different part to test each time. To further increase the accuracy of the precision calculation, this whole process, dividing the learning dataset into ten bins and testing each combination, is repeated ten times. This results in a total of 100 tests for the classifier and learning dataset (Witten et al., 2011).

From the table it can be seen that the precision is on par with the precision obtained in other research. While the precision and recall values are good for gender and age classification, the precision and recall values for the individual classes show problematic values. The large differences in precision and recall values between the individual classes lead to a bias in the output of the classifier. E.g. if the classifier finds females with high precision and males with low precision, both having high recall, than an overestimate of males is expected. The classifier will label a review as belonging to a male more often, and be wrong more often, hence the low precision.

Ten fold cross validation does not say anything about how close to the optimal trained classifier the classifier is. In case of gender classification a baseline will be around 50%, e.g. in picking female every time you are expected to be correct in half the cases. In table 6.6 it is shown that the classifier achieves a precision of 69.4% for gender, but this value provides little information about how much room there is for improvement with respect to the current classification exercise. When there is room for improvement, the problem regarding the classification bias could possibly be solved by having a larger dataset.

Table 6.6.: Precision and recall values for the Multinomial Naïve Bayes classifier using different learning datasets. In the bottom table for age classification, the values between parentheses is for three-class classification using (young, mid, and old) while the other values are for two-class classification, excluding class young.

Gender classification

	blog	**blog functional**	**book review**
precision	61.6%	39.4%	69.4%
recall	97.1%	94.8%	99.3%

Female classification

	blog	**blog functional**	**book review**
precision	68.0%	18.3%	85.8%
recall	97.3%	89.9%	98.6%

Male classification

	blog	**blog functional**	**book review**
precision	46.6%	88.9%	28.8%
recall	96.3%	97.5%	94.3%

Age classification

	blog	**blog functional**	**book review**
precision	39.8% (31.8%)	45.7% (31.8%)	72.8%
recall	95.3% (94.8%)	95.8% (94.8%)	99.4%

Mid classification

	blog	**blog functional**	**book review**
precision	77.1% (49.5%)	58.4% (55.7%)	25.0%
recall	94.7% (92.2%)	92.4% (91.9%)	92.6%

Old classification

	blog	**blog functional**	**book review**
precision	27.5% (26.0%)	41.6% (42.7%)	88.6%
recall	95.9% (96.4%)	97.4% (97.4%)	99.3%

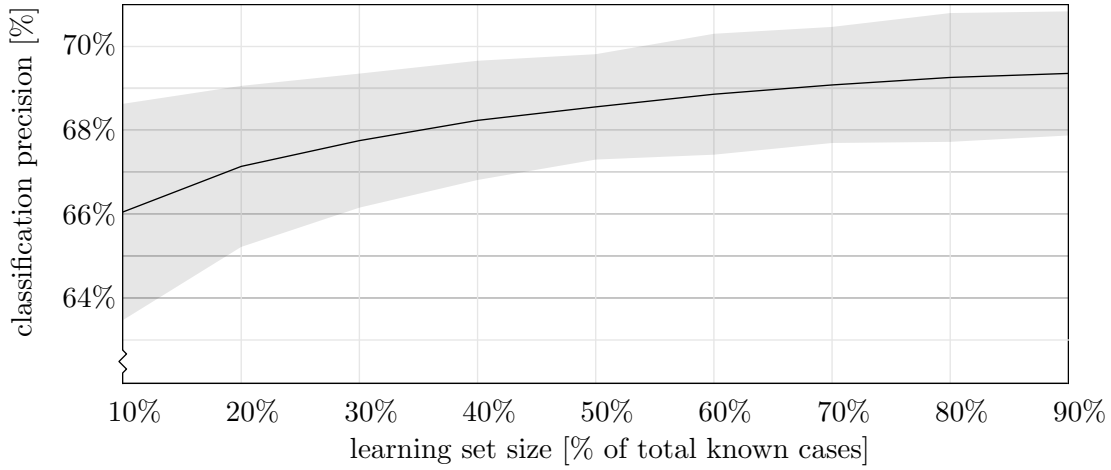


Figure 6.5.: Convergence of the gender classification precision versus the learning set size.

To investigate how good the 69.4% precision is, a convergence plot of the precision versus the size of the learning set is shown in figure 6.5. This precision in this plot is calculated in a similar fashion as the ten fold cross validation. In the Da Vinci Code case there are 8,702 reviews with known author age and gender. These reviews with corresponding meta-data are divided at random into ten parts. For the 10% learning set size only one part is used for training, the remaining nine parts are used to validate the algorithm. For the 20% learning set size two parts are used to learn and the other eight to validate, etcetera. 100 tests have been performed for each point in the graph, using random part divisions between the learning and validation set for the learning set size between 20% and 80%. The standard deviation between these tests is calculated and the 95% confidence bounds are shown in the shaded area in the graph

From figure 6.5 one expects the accuracy to have an upper limit around 70%–71%. Observing the state of convergence, the biases towards a specific class are not expected go away when a larger training set is used. As previously mentioned, these biases introduce a problem in estimating the demographic indicators. For completeness, the results of application of this method on the reviews with missing demographic indicators on the profiles are shown in table 6.7. The large deviation between the different learning sets used is explained by the class bias, making the results unreliable and unusable for this research.

A negative result is given for the different propositions related to using writing style analysis to extract gender and age due to bias in the classifier. The results with respect to the propositions are given in table 6.8. Since the necessary design propositions were not satisfied, the depended empirical propositions cannot be determined from writing style analysis.

Table 6.7.: Absolute numbers for the different classes where gender and age were missing for using different training sets.

training set	blogs	blogs functional	reviews
females	1.818	762	842
males	1.304	2.351	2.600
unknown	581	590	261
mid	7.183	9.304	2.398
old	5.881	3.930	11.651
unknown	1.564	1.394	579

Table 6.8.: Overview of the propositions and results related to writing style analysis.

	Design proposition	**Result**
DP1.6	Writing style analysis can be used to find gender	–
DP1.7	Writing style analysis can be used to find age	–

	Empirical proposition	**Result**
EP1.1	Demographic variables can explain differences in sentiment	?

6.4. Manipulation amongst books over the financial crisis

A collection of reviews from books over the financial crisis is investigated to search for any possible manipulation. This collection of reviews allows for testing the different indicators that can make a review or user account suspicious, but actual classifying a post as manipulated remains subjective. Two indicators, the near-duplicate score and product bias, will be used to find suspicious reviews in the dataset. These selected suspicious reviews will be further analyzed by observing their impact factor and their polarity deviation to help judging whether they are manipulating the Online sentiment or are genuine.

6.4.1. Suspicious reviews from a near-duplicate score

Different researchers have used near-duplicate reviews as an indication for manipulated reviews (see table 4.3). The technique often used is a shingle method (Broder, 2000), and the implementation here is discussed first. Second, near-duplicate reviews results per book are present, followed by near-duplicates across the whole collection of reviews for the books.

Implementation of the shingle method

As discussed in section 4.2.3, near-duplicates can be detected using the shingle method. In the shingle method a review is split up into a feature set, typically a set of n-grams. In case of large collections of reviews that have to be analyzed, a "fingerprint" is used. This is a subset of the n-grams belonging to the review to estimate an initial similarity. From this initial similarity, likely candidates are selected that are further analyzed (Broder, 2000).

Here, the features are 1-grams, i.e. single words. Due to the small collection of reviews, all features are used in the "initial" similarity calculations, making further analyzing unnecessary. The unique features are each given their own dimension and a score of the normalized usage frequency (0 being not used, 1 being the only feature used in the review). In effect, this makes the review a vector of features. The distance between two review-vectors can then be used to calculate a candidate near-duplicate score[5]:

$$s(x, y) = 1 - \sqrt{\frac{\sum_{\text{features}} (x_i - y_i)^2}{N_{\text{features}}}^2}, \tag{6.4}$$

where s is the similarity score, x and y the feature vectors belonging to the document, and N_{features} the amount of unique features in both documents. This equation above will result in a score of 1 for documents that share all features, with the same usage occurrences. The near-duplicate score of a review is the maximum near-duplicate score between the review under investigation and the other reviews in the set.

Near-duplicates on the same book

The method is applied first on the collection of reviews for each book separately. Reviews are marked as near-duplicate when if their near-duplicate score is at least 0.9 (90%). The results are shown in absolute numbers as well as percentages of total reviews per book in table 6.9. All found near-duplicates are from different user accounts and posted on different dates.

For the majority of books in this collection there are no near-duplicates found. In case of the highest ranked book, "The Big Short" by Lewis, there are 9 near-duplicates. Of these 9, 7 are short texts: 3 times "Loved it!", 2 times "Very good!", and 2 times "Should be required reading." (or similar). The other near duplicate texts, also the ones found for "13 Bankers" by Johnson and Kwak and "House of Cards" by Cohan, are longer reviews.

As far as the impact of a review score on the average at time of posting, most are negligible with an influence of less than a percent on the average score. 6 reviews have a larger impact, scoring a 1.0, 0.125, two times 0.025, 0.026, and 0.013. Interestingly, the reviews with highest impact factors are duplicates of each other but have been given different star ratings (0.2 compared to -0.6). The first of these duplicates appears to be

[5]The near-duplicate score is defined as the maximum similarity between the review and all other reviews in the set, hence the usage of candidate here.

Table 6.9.: Overview of the amount of near-duplicate reviews per book in absolute numbers and % of total.

book rank	near-duplicates		book rank	near-duplicates	
1	9	(0.4%)	**11**	0	(0.0%)
2	0	(0.0%)	**12**	0	(0.0%)
3	0	(0.0%)	**13**	0	(0.0%)
4	0	(0.0%)	**14**	0	(0.0%)
5	2	(1.8%)	**15**	0	(0.0%)
6	0	(0.0%)	**16**	0	(0.0%)
7	2	(1.4%)	**17**	0	(0.0%)
8	2	(25.0%)	**18**	0	(0.0%)
9	0	(0.0%)	**19**	0	(0.0%)
10	0	(0.0%)	**20**	0	(0.0%)

a regular user, the second is from a blog on book reviews. All longer reviews appear to be summaries from blogs that were posted for different versions of the book, but merged by Goodreads.com for a single version. The reviews appear to be promotional for the blogs.

Near-duplicates in the whole collection

A manipulator could post similar reviews amongst different books; perhaps a template to promote a single book in the set while being downplaying the sentiment of the other books. If a manipulator applies such a strategy, it will show up when near-duplicates are sought over all reviews for the 20 books in the financial crisis dataset.

Application of the near-duplicate algorithm identified 29 near-duplicates, which corresponds to 0.7% percent of the 3,981 reviews in total. In addition to the duplicates already found for the books separately, 10 others are found. These new found near-duplicate reviews are all short reviews and consist of “Excellent read”, “Great read”, “Highly recommended!”, “A must-read”, another “Loved it!”, and one user issuing visitors to read the review on his/her website.

Furthermore, a histogram showing the maximum similarity scores between reviews is presented in figure 6.6. The figure shows that the majority of reviews share a part of the words with other reviews in the collection. This result is expected as roughly the same topic is discussed in every review, a book about the financial crisis. The peaks at a similarity score of 1.0 and 0.94 are the previously discussed near-duplicates found by this algorithm.

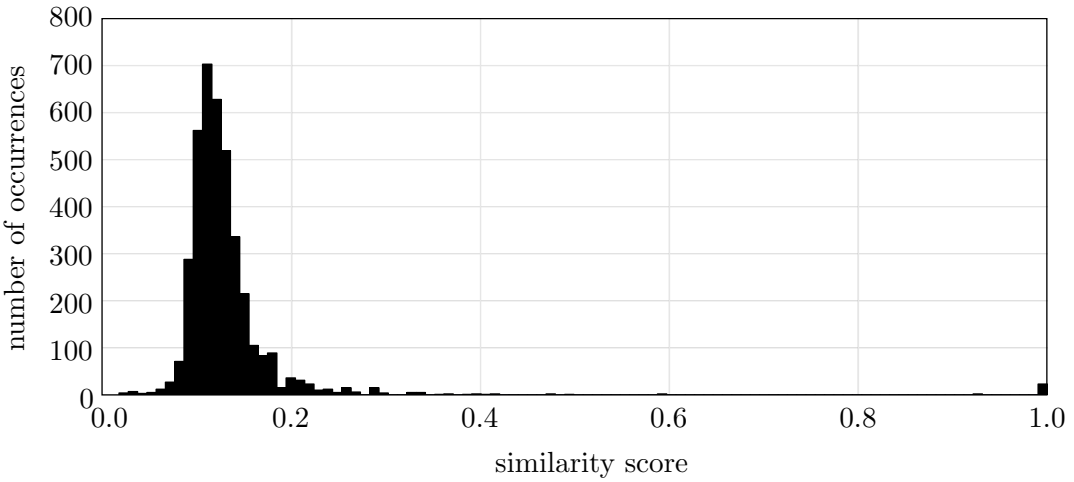


Figure 6.6.: Histogram of the largest similarity score of a review compared with the whole collection of reviews on books over the financial crisis.

Results

The fact that a review has a near-duplicate within the same set is not considered a reason that the review is manipulative. Common phrases and blog summaries/promotions appear to explain all found near-duplicates. Even the more suspicious reviews, due to high impact factor, did not appear to be manipulative. This does not imply that looking for near-duplicates cannot facilitate in detecting manipulation, however, there is no evidence found here that manipulative reviews have a higher chance of coming as near-duplicates of each other (DP2.3). Since no manipulation was detected, no statement can be made about the influence of manipulation on opinion mining results (EP2.1). An overview of the proposition and result is given in table 6.10.

Table 6.10.: Overview of the propositions and result related to the near-duplicate score.

	Design proposition	**Result**
DP2.1	The polarity deviation helps in finding manipulation	?
DP2.2	The impact helps in finding manipulation	?
DP2.3	The near-duplicate score helps in finding manipulation	−/?

	Empirical proposition	**Result**
EP2.1	Manipulation can explain differences in sentiment	?

6.4.2. Product bias for manipulative users

Another hypothesized indicator to mark suspicious reviews is the product bias. In the case of the books over the financial crisis, all books are roughly competitive products as they cover a similar topic. Here a search will be for a bias of an author towards a

specific book, the product bias. First, user accounts that post reviews for different books in the financial crisis collection are identified, candidate manipulation accounts. Next, the polarity deviations of the candidate manipulation accounts is examined to search for a possible bias towards a book. Furthermore, a manual inspection of all reviews posted from the candidate manipulation accounts is done to find any promotion for a specific book.

Multiple posting of users

To start, an overview is created of how users have posted multiple reviews for different books. The results are given in table 6.12. The table shows how multiple postings from an account relate to the books. Per connection, the percentage of the total amount of reviews corresponding to the book of the column is shown. It should be noted that there were no multiple reviews from the same user for the same book in the collection (DP2.4).

From the table it can be seen that for five books, roughly half or more of the reviews come from people that also posted a review for an other financial book (or books). These books are (5) "The Greatest Trade Ever" by Zuckerman, (6) "13 Bankers" by Johnson and Kwak, (7) "The End of Wall Street" by Lowenstein, (12) "The Murder of Lehman Brothers" by Tilman, and (15) "In FED We Trust" by Wessel. No clear manipulation strategies between books directly are seen in the table, the percentages seem to follow the distribution of candidate manipulation accounts versus books.

Product bias and polarity deviation

The next question is whether the users that post reviews for different books in the collection advocate one book in particular. To do this, two indicators are calculated: (1) the polarity deviation of the review compared to the average for the book, and (2) the product bias, i.e. the deviation for all reviews of the candidate manipulative account.

While preforming the calculations, a problem emerged in application of the opinion mining algorithm at the fine-grained level of observing single reviews. People tend to write reviews with a small summary of the book. In this summary sometimes words are used that have meaning to the sentiment classifier, while they do not relate to the opinion of the review author.

An example is the title "The Greatest Trade Ever", which scores full positive on the sentiment scale due to "greatest". When an author uses this book's title in a review it will, therefore, appear more positive due to the sentiment classifier than it is in reality. As Goodreads.com also provides an user-provided star rating, these are used overcome the previously mentioned problem.

Results

For the sentiment score of a review both the opinion mining algorithm and self-provided star score have been tried. Both these metrics revealed no manipulation in combination with both the polarity deviation and product bias. The 892 posts from candidate manipulation accounts are manually inspected and do not reveal any suspicious behavior.

From these reviews, 49 referred to one of the other books in the collection, sometimes recommending one because book as a good companion (22 times), or preferred one of the books (19 times). By far, "The Big Short" takes part in these comparisons (32 times), followed by "Too Big to Fail" (22 times), as expected from the popularity of these books. On a side note, "The Big Short" is often compared with other works of Lewis. Table 6.11 gives an overview of the performance of the method in relation to the design propositions. Furthermore, as again no manipulation is found, no statement can be made about the empirical proposition related to manipulation.

Table 6.11.: Overview of the propositions and findings related to user brand bias.

	Design proposition	**Result**
DP2.4	The reviews per product helps in finding manipulation	?
DP2.5	The product bias helps in finding manipulation	−/?

	Empirical proposition	**Result**
EP2.1	Manipulation can explain differences in sentiment	?

Table 6.12.: Overview of the postings from the same account for different books. The numbers indicate how many users have that have written a review for book X have also written a review for book Y as percentage of the total amount of reviews for the book with the column's rank.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
1	-	16.7	16.0	9.9	20.9	21.2	17.2	16.4	6.2	11.6	2.3	0.0	8.9	6.5	11.5	21.6	9.1	8.0	9.1	0.0
2	4.4	-	8.8	1.4	15.1	6.0	13.7	8.0	0.0	9.8	4.5	0.0	10.4	9.5	13.2	3.7	5.1	3.3	2.4	0.0
3	0.6	1.2	-	0.9	2.0	2.7	4.0	1.2	0.0	5.3	0.0	0.0	0.0	2.4	3.6	1.1	0.0	1.7	0.8	0.0
4	0.3	0.2	0.8	-	0.0	1.5	3.6	1.9	0.0	0.2	0.0	0.0	1.6	1.0	0.9	0.0	1.3	0.4	3.0	0.0
5	0.7	2.0	1.9	0.0	-	2.1	1.6	1.1	0.0	1.3	0.0	0.0	1.6	0.8	1.3	1.9	0.6	0.8	2.3	0.0
6	1.1	1.2	3.8	2.1	3.1	-	3.6	1.5	0.0	2.5	0.0	0.0	0.0	0.0	7.0	2.7	1.5	3.7	1.5	0.0
7	0.3	1.0	2.2	2.0	0.9	1.4	-	2.4	0.0	0.9	0.0	0.0	0.0	1.7	1.0	1.2	0.5	1.7	1.1	0.0
8	1.1	2.0	2.2	3.4	2.0	1.9	7.8	-	0.0	2.9	0.0	0.0	4.7	5.6	4.3	0.7	1.5	0.8	1.5	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0	0.0
10	0.5	1.6	6.2	0.2	1.6	2.1	2.0	1.9	0.0	-	0.0	50.0	1.0	3.2	5.9	0.3	0.2	0.0	0.0	0.0
11	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	0.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.0	-	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.1	0.6	0.0	0.6	0.7	0.0	0.0	1.1	0.0	0.4	2.3	0.0	-	0.0	2.6	0.0	0.3	1.7	0.0	0.0
14	0.1	0.3	0.6	0.3	0.2	0.0	0.9	0.8	0.0	0.7	0.0	0.0	0.0	-	0.5	0.1	0.0	0.0	0.3	0.0
15	0.2	0.9	1.8	0.5	0.7	2.4	0.9	1.2	0.0	2.5	0.0	0.0	3.1	1.0	-	0.9	0.1	0.3	0.3	0.0
16	1.4	0.9	1.9	0.0	3.6	3.3	4.0	0.6	6.2	0.4	0.0	0.0	0.0	0.8	3.1	-	0.1	1.7	2.3	0.0
17	0.7	1.5	0.0	2.9	1.3	2.2	1.9	1.8	0.0	0.4	0.0	0.0	1.6	0.0	0.5	0.2	-	2.5	2.4	14.3
18	0.2	0.3	1.3	0.3	0.7	1.9	2.3	0.4	0.0	0.0	0.0	0.0	3.1	0.0	0.5	0.7	0.9	-	1.9	0.0
19	0.3	0.3	0.6	2.5	2.0	0.9	1.6	0.7	0.0	0.0	0.0	0.0	0.0	1.0	0.5	1.1	0.9	2.1	-	0.0
20	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.6	0.0	0.0	-
%	11.9	30.7	48.1	26.9	54.7	49.6	65.1	40.8	12.5	39.8	9.1	50.0	37.5	33.3	56.4	36.4	22.8	28.8	28.8	14.3
N	264	178	38	21	41	56	28	58	1	37	2	1	12	7	22	51	38	17	19	1

6.5. Events in Drive reviews

The Drive case is used to investigate whether events have an impact on the public sentiment (EP3.1) and whether events can attract a different audience (EP3.2). Two methods are tried to find the launch of the Oscar nominated movie in the reviews, first by looking at changes in post frequencies, second by looking for changes in word-usage frequencies. The feasibility of both methods in event detection is tested by design propositions DP3.1–DP3.2.

6.5.1. Observing changes in post frequencies

Marketing efforts for the film "Drive" or the overall popularity of the film could have a positive effect on the book sales and the amount of reviews. To analyze this property, a plot is made of the post frequencies over time (see figure 6.7). The displayed post frequencies are the amount of reviews per month. Taking the time span of a month allows for smoothing of the data compared to looking at the frequencies between single posts. Furthermore, the time span is expected to be much smaller than the effect of a movie on the reviews, therefore allowing the event to show up in the results.

As can be seen from figure 6.7, about one and a half month after the USA release of the movie Drive, the postings show a dramatic increase. In the years before, about six reviews per year were posted, while after the movie release that same number is easily made within a single month.

A second burst in the amount of reviews is seen starting from January 2012. A possible explanation for this can be reading during the Christmas holidays and posting a review in January. The idea of such seasonal review posting is supported by the posting frequencies of the Da Vinci code case, where spikes can be seen at Januaries throughout the years (see figure 6.7 (b)). Another recurrent spike is seen half-way during the year for both data sets, again suggesting a seasonal effect.

Further investigation of the influence of months by examining the frequency spectrum of the post dates of reviews revealed only a clear peak corresponding with weekly patterns. Even the largest dataset, the Da Vinci Code, is too noisy and has too a too short time span to observe whether a clear peak exists around the frequency for half a year.

6.5.2. Observing changes in word usage frequencies

As proposed by Landmann and Zuell (2007), events can be extracted by monitoring the usage frequencies of words over time. An event influencing a significant amount of reviews can introduce a sudden rise in the usage of the word (or set of words) that is relevant to the event, by the authors naming or describing the event. Hence, deviations in word usage frequencies are investigated for automatic extraction of events.

A method to analyze the words that undergo a large change in usage frequency is given by looking at the derivative of the usage frequency with respect to time,

$$\frac{\partial}{\partial t} f_{word}, \tag{6.5}$$

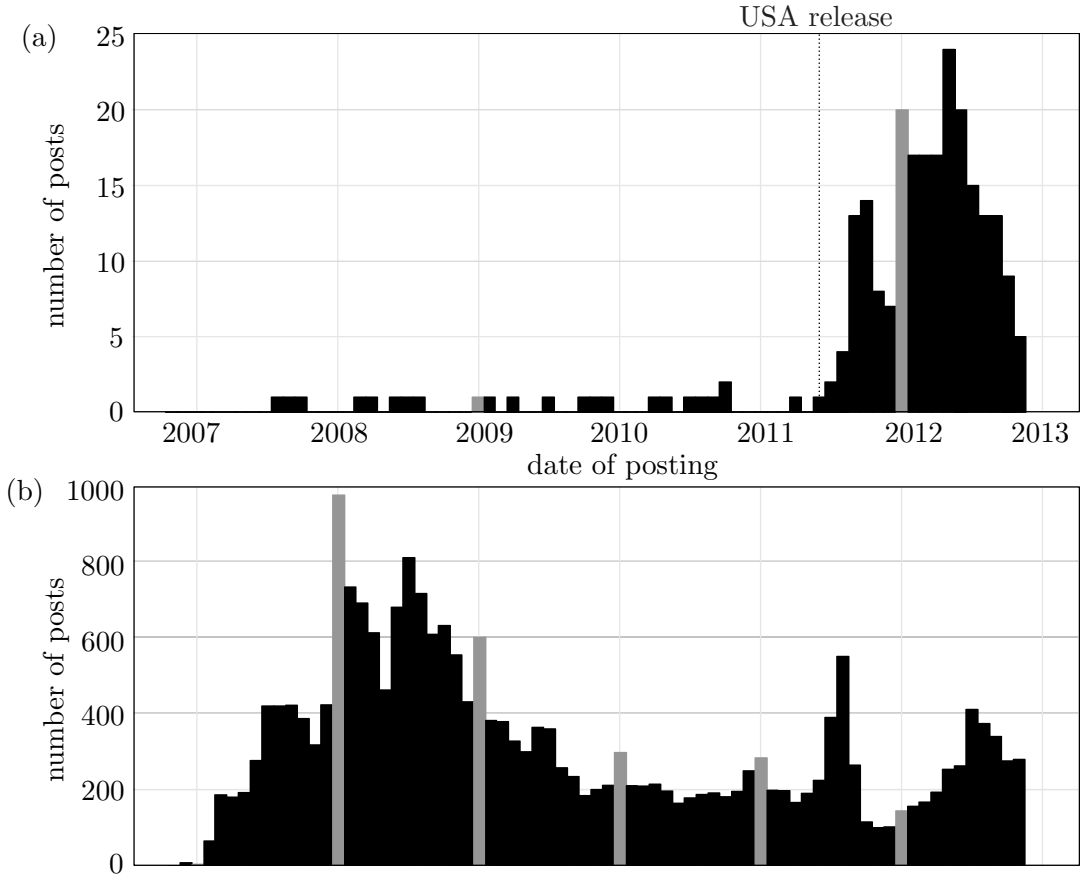


Figure 6.7.: Histogram showing the amounts of post per month versus time for the Drive case (a) and the Da Vinci Code case (b). The dashed black line marks the USA release date of the movie based on the novel, the gray bars indicate January.

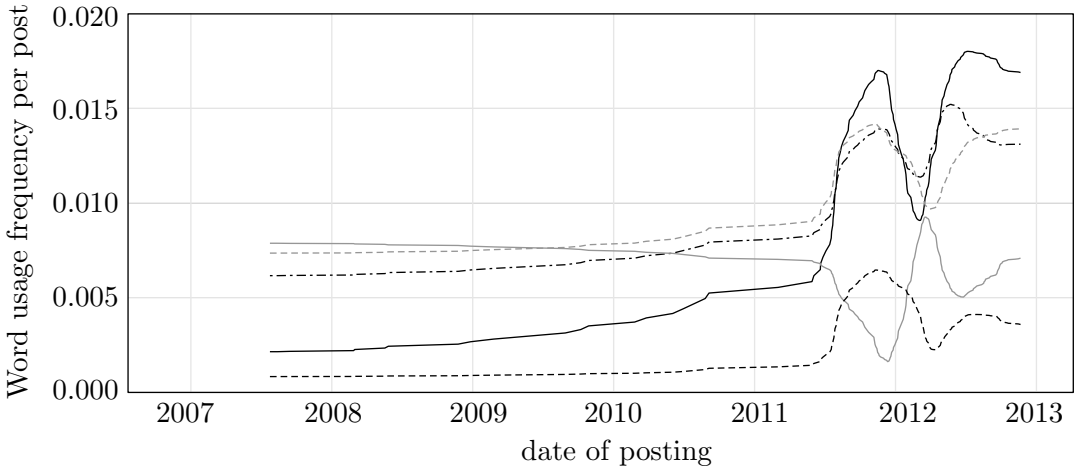


Figure 6.8.: The top five largest changing word usage frequencies over time according to the highest rising slope method: movie (–), book (·-·), noir (–), read (··), different (--)

where f_{word} is the word usage and t time. The word usage frequency data is smoothed over time before analysis, hence, searching the largest slope will result in words that undergo a large and sudden change.

Furthermore, stop words and functional words are excluded to participate in the event finding algorithms. These words are expected to provide little to no information in describing events.

Results

The words with the largest deviations in usage frequency over time are shown in table 6.13. It can be seen that the word "movie" shows up on top, meaning that the usage of this word showed the largest changes over time. The other words that show up could be related to the movie event. Landmann and Zuell (2007) propose as last step in the identification of events that such words should be clustered. Figure 6.8 shows the usage frequencies of the top five words that show the largest deviations. From the similarities in the usage frequencies of the different words over time, it seems well possible that the words "movie", "book", "read", and "different" can indeed be coupled to the same event. "Noir", on the other hand, shows an inverse of the other usage frequencies, perhaps indicating that this word is not often used by moviegoers.

The next questions are whether the event of a movie has an impact on the sentiment (EP3.1) and whether the event attracted a different audience to the book (EP3.2). To investigate this, two different methods that separate the sample into a group influenced by the event and a group not influenced by the event, are tried.

The first method to find the groups is by selecting the influenced group by using word features in their review: "movie", "Gosling" (actor in the movie), and "different". "Noir" is used to identify the group that was not influenced by the movie.

Table 6.13.: words with largest deviations in usage frequency over time for the Drive case

feature	score
movie	1.00
book	0.52
noir	0.59
read	0.46
different	0.48
character	0.37
characters	0.35
American	0.33
short	0.32
time	0.31

Table 6.14.: Comparison between authors influenced by the movie (inf.) and authors not influenced by the movie (not-inf.) using two separation methods.

	words		**time**	
	inf.	**not-inf.**	**inf.**	**not-inf.**
group size	151	113	45	23
average sentiment	0.20	0.13	0.15	0.12
average stars	0.30	0.24	0.14	0.17
male % (males)	59 (75)	71 (67)	62 (26)	90 (19)
average age (known values)	34 (64)	39 (43)	33 (22)	55 (9)

The second method divides the sample into two groups by looking at the post date. The movie premiered in the U.S. on June 17, 2011, all reviews before that date are collected in the group of not influenced, the reviews posted over six months from that date are collected into the influenced group. Doing so captures the first peek in postings seen in figure 6.7 (a). A comparison of both groups created by the different methods is given in table 6.14.

While the small group sizes make drawing conclusions difficult, it is interesting to see that the same characteristics of the event are shown by both separation methods. In case of the influenced group, the sentiment from analyzing the review text is more positive compared to the not influenced group. The amount of stars given shows, however, opposite behavior. The influenced group gave lower rating compared to the not influenced group. An explanation for this can be positive feelings expressed towards to movie, but not towards the book. The sentiment classifier used here does not take this into account.

Another explanation can be a different audience. The influenced group shows a higher ratio of females compared to the not influenced group. Furthermore, the not influenced group appears to have a higher age than the influenced group. As mentioned before, the amount of numbers that is averaged over is, however, low and drawing hard conclu-

sions from this data is doubtful. Yet, the correspondence between the different selection methods indicates that the event can indeed attract a different population.

Both methods appear to be promising in identifying events (DP3.1 and DP3.3). Due to the agreement between the indicators one could argue that the event selection methods can find influenced reviews (DP3.2 and DP3.4). These points are covered in the design propositions related to events and give the opportunity to test the three empirical propositions defined for events. The results regarding the propositions are given in table 6.15.

Table 6.15.: Overview of the propositions and results related to events.

	Design proposition	**Result**
DP3.1	Word-frequencies can identify an event	+
DP3.2	Word features can be used to find influenced reviews	+ / ?
DP3.3	Post-frequencies can identify an event	+
DP3.4	Date and time of posting can be used to find influenced reviews	+ / ?

	Empirical proposition	**Result**
EP3.1	Events can explain differences in sentiment	+ / ?
EP3.2	Events can attract a different population	+

6.6. Results overview

Mixed results have been found related to the empirical and design propositions, an overview of all propositions is given in table 6.16. Not all proposed methods could confirm or disprove the influence of the different bias sources. Yet, in some cases a substitution for a different method could be used.

Table 6.16.: Overview of all propositions and the results of this research.

	Design proposition	**Result**
DP1.1	Profile extraction can be used to find gender	+
DP1.2	Profile extraction can be used to find age	+
DP1.3	Profile extraction can be used to find location	+
DP1.4	Profile extraction can be used to find education	-
DP1.5	(User)name analysis can be used to find gender	+
DP1.6	Writing style analysis can be used to find gender	−
DP1.7	Writing style analysis can be used to find age	−
DP2.1	The polarity deviation helps in finding manipulation	?
DP2.2	The impact helps in finding manipulation	?
DP2.3	The near-duplicate score helps in finding manipulation	−/?
DP2.4	The reviews per product helps in finding manipulation	?
DP2.5	The product bias helps in finding manipulation	−/?
DP3.1	Word-frequencies can identify an event	+
DP3.2	Word features can be used to find influenced reviews	+ / ?
DP3.3	Post-frequencies can identify an event	+
DP3.4	Date and time of posting can be used to find influenced reviews	+ / ?

	Empirical proposition	**Result**
EP1.1	Demographic variables can explain differences in sentiment	+
EP2.1	Manipulation can explain differences in sentiment	?
EP3.1	Events can explain differences in sentiment	+ / ?
EP3.2	Events can attract a different population	+

7. Discussion and conclusions

This research discussed problems with the external validity of opinion mining research. Application of external validity theory to the field of opinion mining revealed three forms of possible biases, which were investigated here: (1) a mismatch in demographic indicators of the sample and target population, (2) opinions influenced by events, and (3) manipulation of online reviews.

A literature study is performed to find descriptive indicators for these bias forms and various methods to extract the indicators from Online content. Different cases are selected through theoretical sampling to test whether these problems occur in opinion mining and should be taken into account by researchers. All cases cover book reviews and are obtained from Goodreads.com.

7.1. Influence of demographic variables

To investigate the importance of gender and age on the opinion mining results, all reviews from Goodreads for "The Da Vinci Code" by Dan brown were collected. Using the self-provided information of review authors, it is shown that females overrepresent the sample, approximately 74% females versus 26% males. Furthermore, as the reviewers get older, the differences in sentiment between males and females become more prominent. Females in their 40s are on average 30% more positive than the males in that age group. The same effect is seen when self-provided ratings given by the authors are used instead of sentiment scores based on the review texts, supporting the existence of demographic biases in opinion mining research.

As the profile data not always includes the demographic variables of interest, two methods were tested to collect the gender and age of reviewers. The first method estimates gender by looking at the name of a review author, the second method estimates gender and age from the writing style. Estimating of gender from the user's name shows highly accurate results (99.3% and 98.8%) and is able to recover large parts of the test set (80.5%). While reasonable accuracies are also obtained by application of writing style analysis (69.4% for gender, 72.8% for age), the method has a bias towards a specific class. This bias makes application of writing style analysis not suitable to recover missing ages and genders in this research.

7.2. Manipulation of reviews

Reviews for a collection of 20 books covering the financial crisis of 2008 are used to find the effect of manipulated reviews in opinion mining research. By looking at near-

duplicates, often performed in literature to identify manipulation, suspicious reviews were selected. The shorter near-duplicates appear to be common sayings as "Loved it!". Whether these text are posted with the intention to manipulate is doubtful, as the impact of these reviews on the average rating at the time of posting is often negligible in the presented case ($< 1\%$). Other near-duplicates could be traced back to be summaries from blogs on books and appeared to be more promotional for the blog than manipulating Online sentiment. Near-duplicates do not reveal manipulation in this research.

A second approach to find manipulation is tried by searching for users with a strong preference for a specific book. Users that posted for multiple books in our financial crisis set were selected. From these accounts a bias was sought were they would promote one book while downplaying the sentiment of competing books. Both analysis from the sentiment as well as manual inspection of these reviews reveals no manipulation amongst the reviews.

7.3. Influence of events

Two methods are used to monitor the effects of events on the Online sentiment. Book reviews for the book "Drive" by James Sallis are chosen as the book is used for a (popular) movie. By observing changes in the posting frequency one can successfully identify a large increase shortly after the movie's premiere. Furthermore, using a clustering approach by looking at changes in word usage frequencies over time, the movie event is successfully discovered and described by words as "movie", "read", "book", and "different".

The event appears to modify the population that posted reviews. The reviewers influenced by the movie-event appear younger and are more often female. The effect of the event with respect to the Online sentiment is uncertain. By selecting influenced reviews based on the appearance of certain words in the review, a large difference in sentiment between the influenced and not influenced reviews is seen; influenced reviews are about 50% more positive in sentiment score and 30% more positive in self-provided rating. Selecting influenced reviews by a time span did, however, not reveal such dramatic difference in sentiment score (approximately 20%) and a reversed relation in self-provided rating (approximately -30%).

7.4. Limitations of the research

The results presented here are based on self-provided user data. This data can be false and introduce errors in the demographic data used throughout this research. Likewise, whether a review for Drive is actually influenced by the movie or not, and whether a review is posted to manipulated or not, is never known for sure.

Furthermore, the Goodreads API only allowed for collecting approximately 75% of all the reviews available per book. In the API documentation they describe that the most popular reviews are returned. Their algorithm for determining the popularity of a review is unknown but could introduce a bias in the presented results. One could

expect manipulative reviews to be less popular, and this can be an explanation why no manipulated reviews are found in this research.

7.5. Recommendations for practice

The research showed that demographic biases indeed occur in opinion mining research. With respect to gender, the Da Vinci Code case showed an overrepresentation of females while for the Drive case males were the dominant group. Furthermore, differences in sentiment between gender and age were seen. Opinion mining researchers should be aware of these effects, and show how demographic variables influence their findings.

Incorporating the author context model in opinion mining research allows researchers to test public sentiment more thoroughly. For example, corporations could identify problematic adoption of their product by certain target groups or test if a targeted campaign indeed reaches their target and has the desired sentiment effect. Perhaps using the author context model can even facilitate in identifying new niches.

7.6. Suggestions for future research

For future research I recommend picking a much discussed product with new versions coming out over the years. This will allow for more comprehensive testing of the influence of events. When doing so, an opinion mining algorithm that can extract sentiment related to features should be used. This creates a clear division between sentiment related to the product and other items.

Different websites should be used to search for possible differences in sentiment with respect to the website setting. Such a study can be compared with a multi-case study in social sciences and can strengthen the findings of this thesis.

With respect to opinion mining algorithms, the results that show comparisons between the ratings given by the reviewer and estimate sentiment, the average estimated sentiment is often lower than the rating. I expect this to be due to short texts, foreign texts, and irrelevant sentiment features matches. Analyzing the text in more detail using natural language processing techniques could help overcome these problems.

Furthermore, other algorithms proposed in the meta-design but not further used here can still be tested. The implementation of the methods might be cumbersome, but extracting demographic indicators from profile photos could prove to be valuable. One might even try to extract ethnicity from the profile photos and family names, searching for interesting different demographic biases.

8. Bibliography

Abel, F., N. Henze, E. Herder, and D. Krause (2010). "Interweaving public user profiles on the web". In: *User Modeling, Adaptation, and Personalization, 18th International Conference*, pp. 16–27.

Abe, S., K. Inui, K. Hara, H. Morita, C. Sao, M. Eguchi, A. Sumita, K. Murakami, and S. Matsuyoshi (2011). "Mining personal experiences and opinions from Web documents". In: *Web Intelligence and Agent Systems: An International Journal* 9, pp. 109–121.

Adomavicius, G. and A. Tuzhilin (1999). "User Profiling in Personalization Applications through Rule Discovery and Validation". In: *International Conference on Knowledge Discovery and Data Mining*, pp. 377–381.

Argamon, S., M. Koppel, J. Fine, and A. R. Shimoni (2003). "Gender, genre, and writing style in formal written texts". In: 23, pp. 321–346.

Argamon, S., M. Koppel, J. W. Pennebaker, and J. Schler (2009). "Automatically profiling the author of an anonymous text". In: *Communications of the ACM* 52.2, pp. 119–123.

Babbie, E. R. (2007). *The Practice of Social Research*. 11th ed. Cengage Learning.

Balaguer, E. V. and P. Rosso (2011). "Detection of Near-duplicate User Generated Contents: The SMS Spam Collection". In: *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pp. 27–33.

Balduzzi, M., C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel (2010). "Abusing social networks for automated user profiling". In: *Proceedings of the 13th international conference on Recent advances in intrusion detection*, pp. 422–441.

Becker, H., M. Naaman, and L. Gravano (2010). "Learning similarity metrics for event identification in social media". In: *Conference on Web search and data mining*, pp. 291–300.

Broder, A. Z. (2000). "Identifying and filtering near-duplicate documents". In: *Combinatorial Pattern Matching*. Ed. by R. Giancarlo and D. Sankoff. Springer-Verlag Berlin / Heidelberg, pp. 1–10.

Brun, C. (2011). "Detecting Opinions Using Deep Syntactic Analysis". In: *Proceedings of Recent Advances in Natural Language Processing*. September, pp. 392–398.

Can, F. and J. M. Patton (2004). "Change of Writing Style with Time". In: *Computers and the Humanities* 38.1, pp. 61–82.

Caverlee, J. and S. Webb (2008). "A large-scale study of MySpace: Observations and implications for online social networks". In: *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pp. 104–114.

Chandramouli, R. and K. P. Subbalakshmi (2009). "Gender identification from E-mails". In: *2009 IEEE Symposium on Computational Intelligence and Data Mining*, pp. 154–158.

Chandy, R. and H. Gu (2012). "Identifying spam in the iOS app store". In: *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality - WebQuality '12*, p. 56.

Cheng, N., R. Chandramouli, and K. Subbalakshmi (2011). "Author gender identification from text". In: *Digital Investigation* 8.1, pp. 78–88.

Dahllof, M. (2012). "Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches–A comparative study of classifiability". In: *Literary and Linguistic Computing* 27.2, pp. 139–153.

Das, A. and S. Bandyopadhyay (2011). "Towards the Global SentiWordNet". In: *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pp. 799–808.

Das, A., S. Bandyopadhyay, and B. Gambäck (2012). "Sentiment Analysis: What is the End User's Requirement?" In: *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*.

Dave, K., S. Lawrence, and D. M. Pennock (2003). "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews". In: *Proceedings of WWW*, pp. 519–528.

Dellarocas, C. (2006). "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms". In: *Management Science* 52.10, pp. 1577–1593.

Dijck, J. van (2009). "Users like you? Theorizing agency in user-generated content". In: *Media, Culture & Society* 31.1, pp. 41–58.

Ding, X., B. Liu, and P. S. Yu (2008). "A Holistic Lexicon-Based Approach to Opinion Mining". In: *Proceedings of the international conference on Web search and web data mining*, pp. 231–239.

Estival, D., T. Gaustad, B. Hutchinson, S. B. Pham, and W. Radford (2008). "Author Profiling for English and Arabic Emails". In: *Natural Language Engineering*, pp. 1–22.

Fawcett, T. and F. Provost (1996). "Combining Data Mining and Machine Learning for Effective User Profiling". In: *International Conference on Knowledge Discovery and Data Mining*, pp. 8–13.

Filippova, K. (2012). "User Demographics and Language in an Implicit Social Network". In: *Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning*. July, pp. 1478–1488.

Fukuhara, T., H. Nakagawa, and T. Nishida (2007). "Understanding sentiment of people from news articles: Temporal sentiment analysis of social events". In: *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Gayo-Avello, D. (2012). "A meta-analysis of state-of-the-art electoral prediction from Twitter data". In: *Arxiv preprint arXiv:1206.5851*.

Gayo-Avello, D., P. T. Metaxas, and E. Mustafaraj (2011). "Limits of electoral predictions using twitter". In: *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 490–493.

Geng, X., Z.-H. Zhou, and K. Smith-Miles (2007). "Automatic age estimation based on facial aging patterns." In: *IEEE transactions on pattern analysis and machine intelligence* 29.12, pp. 2234–40.

Gill, A. J. and R. M. French (2007). "Level of representation and semantic distance: Rating author personality from texts". In: *Proceedings of the Second European Cognitive Science Conference*.

Goswami, S., S. Sarkar, and M. Rustagi (2009). "Stylometric analysis of bloggers' age and gender". In: *Proceedings of the Third International ICWSM Conference*. January 1999, pp. 214–217.

Greenberg, S. (2001). "Context as a Dynamic Construct". In: *Human-Computer Interaction* 16.2, pp. 257–268.

Hamilton, R. J. and B. J. Bowers (2006). "Internet recruitment and e-mail interviews in qualitative studies." In: *Qualitative health research* 16.6, pp. 821–835.

Hu, N., I. Bose, Y. Gao, and L. Liu (2011). "Manipulation in digital word-of-mouth: A reality check for book reviews". In: *Decision Support Systems* 50.3, pp. 627–635.

Hu, N., I. Bose, N. S. Koh, and L. Liu (2012). "Manipulation of online reviews: An analysis of ratings, readability, and sentiments". In: *Decision Support Systems* 52.3, pp. 674–684.

Inui, K., S. Abe, K. Hara, H. Morita, C. Sao, M. Eguchi, A. Sumida, K. Murakami, and S. Matsuyoshi (2008). "Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents". In: *International Conference on Web Intelligence and Intelligent Agent Technology*. Ieee, pp. 314–321.

Jindal, N. and B. Liu (2008). "Opinion spam and analysis". In: *Proceedings of the international conference on Web search and web data mining*, pp. 219–229.

Koppel, M. and J. Schler (2006). "The Importance of Neutral Examples for Learning Sentiment". In: *Computational Intelligence* 22.2, pp. 100–109.

Ku, L.-W., T. Liang, and H.-H. Chen (2006). "Opinion extraction, summarization and tracking in news and blog corpora". In: *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, pp. 100–107.

Landmann, J. and C. Zuell (2007). "Identifying Events Using Computer-Assisted Text Analysis". In: *Social Science Computer Review* 26.4, pp. 483–497.

Lanitis, A., C. Draganova, and C. Christodoulou (2004). "Comparing different classifiers for automatic age estimation." In: *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society* 34.1, pp. 621–8.

Lim, P., V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw (2010). "Detecting product review spammers using rating behaviors". In: *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, p. 939.

Li, W. (2012). "Analyses of baby name popularity distribution in US for the last 131 years". In: *Complexity* 48.1, pp. 1–12.

Longueville, B. D., R. S. Smith, and G. Luraschi (2009). ""Omg, from here, i can see the flames!": a use case of mining location based social networks to acquire spatio-temporal data on forest fires". In: *Proceedings of the 2009 International Workshop on Location Based Social Networks*. c, pp. 73–80.

Lu, Y., P. Tsaparas, A. Ntoulas, and L. Polanyi (2010). "Exploiting social context for review quality prediction". In: *Proceedings of the 19th international conference on World wide web - WWW '10*, p. 691.

Mei, Q. and C. Zhai (2005). "Discovering evolutionary theme patterns from text: an exploration of temporal text mining". In: *International Conference on Knowledge Discovery and Data Mining*, pp. 198–207.

Meyerson, P. and W. W. Tryon (2003). "Validating internet research: a test of the psychometric equivalence of internet and in-person samples." In: *Behavior research methods, instruments, & computers: a journal of the Psychonomic Society, Inc* 35.4, pp. 614–620.

Miao, Q., Q. Li, and R. Dai (2009). "AMAZING: A sentiment mining and retrieval system". In: *Expert Systems with Applications* 36.3, pp. 7192–7198.

Min, H.-J. and J. C. Park (2012). "Identifying helpful reviews based on customer's mentions about experiences". In: *Expert Systems with Applications* 39.15, pp. 11830–11838.

Mislove, A., S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist (2011). "Understanding the Demographics of Twitter Users". In: *Fifth International AAAI Conference on Weblogs and Social Media*, pp. 17–21.

Missen, M. M. S., M. Boughanem, and G. Cabanac (2010). "Opinion Detection in Blogs: What Is Still Missing?" In: *2010 International Conference on Advances in Social Networks Analysis and Mining*. Ieee, pp. 270–275.

Morales, F. (2013). *Function Words*.

Mukherjee, A., B. Liu, and N. Glance (2012). "Spotting Fake reviewer groups in consumer reviews". In: *World Wide Web*, pp. 191–200.

Narayanan, A., H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song (2012). "On the Feasibility of Internet-Scale Author Identification". In: *IEEE Symposium on Security and Privacy*.

Narver, J. C. and S. F. Slater (1990). "The effect of a market orientation on business profitability". In: *The Journal of Marketing* October, pp. 20–36.

Oberlander, J. and S. Nowson (2006). "Whose thumb is it anyway? Classifying author personality from weblog text". In: *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. July, pp. 627–634.

Pang, B. and L. Lee (2008). "Opinion Mining and Sentiment Analysis". In: *Foundations and Trends® in Information Retrieval* 1.1-2, pp. 1–135.

Pang, B., L. Lee, and S. Vaithyanathan (2002). "Thumbs up? Sentiment classification using machine learning techniques". In: *Proceedings of the Conference on Empirical Methods in Natural*. July, pp. 79–86.

Park, K. C., Y. Jeong, and S. H. Myaeng (2010). "Detecting experiences from weblogs". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. July, pp. 1464–1472.

Peersman, C., W. Daelemans, and L. Van Vaerenbergh (2011). "Predicting age and gender in online social networks". In: *Proceedings of the 3rd international workshop on Search and mining user-generated contents - SMUC '11*, p. 37.

Perito, D., C. Castelluccia, M. A. Kaafar, and P. Manils (2011). "How Unique and Traceable Are Usernames?" In: *Privacy Enhancing Technologies*. Springer-Verlag, pp. 1–17.

Pfeil, U., R. Arjan, and P. Zaphiris (2009). "Age differences in online social networking – A study of user profiles and the social capital divide among teenagers and older users in MySpace". In: *Computers in Human Behavior* 25.3, pp. 643–654.

Prasath, R. R. (2010). "Learning age and gender using co-occurrence of non-dictionary words from stylistic variations". In: *RSCTC'10 Proceedings of the 7th international conference on Rough sets and current trends in computing*, pp. 544–550.

Ross, M. W., S.-A. Månsson, K. Daneback, A. Cooper, and R. Tikkanen (2005). "Biases in internet sexual health samples: comparison of an internet sexuality survey and a national sexual health survey in Sweden." In: *Social science & medicine (1982)* 61.1, pp. 245–52.

Sarawgi, R., K. Gajulapalli, and Y. Choi (2011). "Gender attribution: tracing stylometric evidence beyond topic and genre". In: *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. June, pp. 78–86.

Saurí, R. and J. Pustejovsky (2012). "Are You Sure That This Happened? Assessing the Factuality Degree of Events in Text". In: *Computational Linguistics* November 2011, uncorrected proof.

Schler, J., K. Mosche, S. Argamon, and J. W. Pennebaker (2006). "Effects of Age and Gender on Blogging". In: *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Shadish, W. R., T. D. Cook, and D. T. Campbell (2002). *Experimental and Quasi-Experimental Designs*. Houghton Mifflin Company.

Singh, R. R. and D. S. Tomar (2009). "Approaches for user profile Investigation in Orkut Social Network". In: *International Journal of Computer Science and Information Security* 6.2, pp. 259–268.

Smedt, T. D. and W. Daelemans (2012). "Pattern for Python". In: *The Journal of Machine Learning ...* 13, pp. 2063–2067.

Social Security Administration, U. S. (2012). *Beyond the Top 1000 Names, National data*.

Sokolova, M. and G. Lapalme (2011). "Learning opinions in user-generated web content". In: *Natural Language Engineering* 17.4, pp. 541–567.

Stone, B. and M. Richtel (2007). *The Hand That Controls the Sock Puppet Could Get Slapped*. (Visited on 06/06/2012).

Teo, T. S. H. and V. K. G. Lim (2000). "Gender differences in internet usage and task preferences". In: *Behaviour & Information Technology* 19.4, pp. 283–295.

Thelwall, M. (2008). "Social networks, gender, and friending: An analysis of MySpace member profiles". In: *Journal of the American Society for Information Science and Technology* 59.8, pp. 1321–1330.

Thelwall, M., D. Wilkinson, and S. Uppal (2009). "Data mining emotion in social network communication: Gender differences in MySpace". In: *Journal of the American Society for Information Science and Technology* 61.1, pp. 190–199.

Tsolmon, B., A.-R. Kwon, and K.-S. Lee (2012). "Extracting Social Events Based on Timeline and Sentiment Analysis in Twitter Corpus". In: *Natural Language Processing and Information Systems*. Springer Berlin / Heidelberg, pp. 265–270.

Turney, P. D. (2002). "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. July, pp. 417–424.

Walls, J. G., G. R. Widmeyer, and O. A. E. Sawy (1992). "Building an information system design theory for vigilant EIS". In: *Information Systems Research* 3.1, pp. 36–59.

Wang, B. and Y.-m. Xue (2011). "Can We Believe the Fund Reviews in Internet Forums?" In: *2011 International Conference on Computational and Information Sciences*, pp. 418–420.

Warren Liao, T. (2005). "Clustering of time series data—a survey". In: *Pattern Recognition* 38.11, pp. 1857–1874.

Witten, I. H., E. Frank, and M. A. Hall (2011). *Data Mining: Practical machine learning tools and techniques*.

Wu, Y., F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu (2010). "OpinionSeer: interactive visualization of hotel customer feedback." In: *IEEE transactions on visualization and computer graphics* 16.6, pp. 1109–1118.

Xie, S., G. Wang, S. Lin, and P. S. Yu (2012). "Review spam detection via temporal pattern discovery". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*, p. 823.

Xu, K., S. S. Liao, J. Li, and Y. Song (2011). "Mining comparative opinions from customer reviews for Competitive Intelligence". In: *Decision Support Systems* 50.4, pp. 743–754.

Ye, Q., Z. Zhang, and R. Law (2009). "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches". In: *Expert Systems with Applications* 36.3, pp. 6527–6535.

Zhao, B., Z. Zhang, Y. Gu, X. Gong, W. Qian, and A. Zhou (2011). "Discovering Collective Viewpoints on Micro-blogging Events Based on Community and Temporal Aspects". In: *Advanced Data Mining*. Springer Berlin / Heidelberg, pp. 270–284.

Ziegler, C.-N. and M. Skubacz (2006). "Towards automated reputation and brand monitoring on the web". In: *IEEE/WIC/ACM International Conference on Web Intelligence, 2006*, pp. 1066–1072.

A. Name-gender combinations for name analysis

A.1. Application in Mislove et al. (2011)

The idea of using names to determine the gender of an user was applied by Mislove et al. (2011). In their work, they collected name and gender information from the U.S. Social Security Administration for the years 1900-2009. An online interface allows querying the database of the SSA for the 1000 most popular male and female baby names in a given year. In case a name appears to be used for both genders, the name is only picked when it is at least 95% predictive in determining the gender.

In order to apply the gender classification, the name is matched with their known name-gender combinations. Mislove et al. (2011) find the user's name from the self-reported name of the users in their Twitter profiles. From this self reported name, the first word is picked and queried in the name-gender database to search the corresponding gender. Using this algorithm, they were able to predict the gender of 64.2% people in their sample by looking for exact matches.

They note three limitations to their gender classification approach. First, user can misrepresent their name. Second, the possibility of different behavior in publishing one's first name compared to gender. Third, they note that the name-gender database may cover different fractions of the male and female populations.

A.2. Training the name-gender classifier

The name-gender relations used in the classification process to determine the gender that corresponds for a specific name are collected from the U.S. SSA (Social Security Administration, 2012). A problem occurs when a name can be used for both males as females. In this section describes three different methods to overcome this problem and select a set of name-gender relations to be used for classification.

The dataset used by Mislove et al. (2011) was build up from selecting the top 1000 most popular name-gender combinations from the SSA over the years 1900–2009. This name-gender dataset is then filtered for names being 95% predictive in gender to come up with the name-gender results used for classification. While they do not give their definition of a name being predictive in gender, here it is thought that being predictive in gender means that the name in combination with one of the genders occurs at least in α times of the total amount of babies given that name, i.e.

$$\frac{N_{n,g}}{N_{n,f} + N_{n,m}} \geq \alpha, \tag{A.1}$$

where $n_{n,g}$ are the amount of babies with a specific name for gender g, f being female and m male. This definition assures that every name corresponds to either the female or male class with an accuracy of α. However, with this definition in prediction accuracy, the resulting amount of name-gender relations (5,386) of Mislove et al. (2011) seems overly optimistic.

The problem lies in the unknown values. For example, a female name appears on the top 1000 list with given number of occurrences. If the male name is not present in the top 1000 most popular names of the same year, then the only thing known about the number of male babies given that same name is that it lies within the range of 0–N, where N is the lowest number of male babies given given a particular name on the top 1000 list, the value at the 1000^{th} place. In other words, the amount of babies with a specific name and gender can be written as

$$N_{n,g} = \sum_{\text{years}} n(y,n,g) \quad \text{with} \quad n(y,n,g) = \begin{cases} d(y,n,g) \text{ if available,} \\ {}^1\!/_2(1 \pm 1)\min(d(y,g)) \text{ otherwise,} \end{cases} \tag{A.2}$$

where d describes the available dataset, $d(y,n,g)$ is the number of babies in the given year y with the given name n and gender g, and $d(y,g)$ is a list of all values in the dataset for a given year and gender.

The plus-minus symbol in the function $n(y,n,g)$ of (A.2) introduces a range of possible values for the amount of year, name, and gender combinations. In order to assure the α accuracy, (A.1) has to be rewritten to

$$\frac{\min(N_{n,g})}{\min(N_{n,g}) + \max(N_{n,\text{other}g})} \geq \alpha, \tag{A.3}$$

which assumes the worst case for each name-gender pair.

If now a similar dataset is downloaded from the SSA as described in Mislove et al. (2011) and an $\alpha = 0.95$ is imposed, the size of the name-gender combinations dramatically decreases. Of the initial 6,178 unique names in the top 1000 most popular baby names over years 1900–2011, only 379 are known to satisfy imposed the 95% accuracy.

A.2.1. Strict 95% precise name-gender set

Luckily, the SSA provides a much larger dataset containing the name, gender, and number of babies with that gender and name born in a year (Social Security Administration, 2012). The dataset contains all records where the name and gender combination have at least 5 occurrences. For (A.2) this truncation of the dataset means that for all name and gender combinations that do not appear have 0 to 4 occurrences.

The dataset consists of 89,873 unique names, which, after enforcing 95% accurate gender prediction from the name results in a set of 2,069 name-gender combinations. The restriction used for creating the 95% accurate set is shown in A.3 by choosing an α of 0.95. Enforcement of this restriction reduces the initial name-gender dataset from 89,873 names to only 2,069 name-gender combinations to be used for the classification algorithm.

A.2.2. Loose 95% precise name-gender set

Assuring 95% precision has a seviere impact on the size of the name-gender database. The size of the is only 2.3% of the initial set. This gives the set a small coverage of names of the U.S. population, see figure A.2. To improve the size of the result set, the restriction imposed by (A.3) can be loosened by using a likely value to replace the unknown values instead of implementing a worst-case scenario. When doing this, an average 95% or better precision is *expected* over the whole dataset but not *guaranteed* for individual names.

To find the likely value of missing data the distribution of the data has to be known. In figure A.1 it is shown that the distribution of the amount of times a similar name occurs per name and year roughly follows a power-law near as proposed by Li (2012). The power law is fitted near the origin using MATLAB's `polyfit` function over the data until 1,100 similar names per name and year. The found power law is described by

$$dist(n) \approx e^{15.16} n^{-1.76}. \tag{A.4}$$

This guess for the distribution of the missing data can now be used to calculate a likely value of the missing data, which yields

$$n_{\text{missing}} = \frac{\sum_{n=1}^{4} n\, dist(n)}{\sum_{n=1}^{4} dist(n)} \approx 1.55. \tag{A.5}$$

Using this value, the $n(y, n, g)$ function of (A.2) is redefined as

$$n(y, n, g) = \begin{cases} d(y, n, g) \text{ if available,} \\ n_{\text{missing}} \text{ otherwise.} \end{cases} \tag{A.6}$$

Application of this equation in combination with the accuracy restriction as defined in (A.1) results in an increased size of the name-gender combinations. The total amount of name-gender combinations is 3,623, slightly above 4% of the initial 89,873 names. Again, the coverage of the U.S. baby population per year is shown in figure A.2.

A.2.3. 95% accuracy while ignoring unknown values

As the previously obtained amount of name-gender relations appears small, another option of dealing with unknowns is presented. In this third method the unknown values are dealt with by ignoring them.

Setting all unknown values to zero,

$$n(y, n, g) = \begin{cases} d(y, n, g) \text{ if available,} \\ 0 \text{ otherwise,} \end{cases} \tag{A.7}$$

results in 84,404 name-gender combinations but the actual accuracy is uncertain. Again, the coverage of U.S. babies throughout the years is shown in figure A.2. The next subsection will describe the coverage in more detail as Mislove et al. (2011) mentioned justly that difference in population coverage per gender can introduce a bias in the classification algorithm.

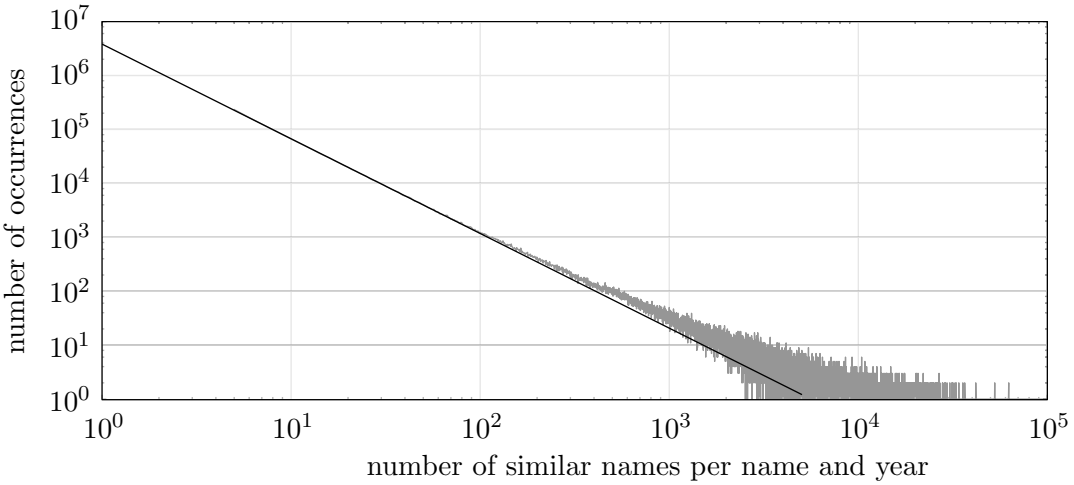


Figure A.1.: Distribution of the amount of similar names per name and year of the SSA dataset in years 1900–2011 in gray and a power law fit near the origin in black.

Table A.1.: Top five U.S. baby names in 2011 from the Social Security Administration.

Male name	**Number of males**	**Percent of total males**	**Female name**	**Number of females**	**Percent of total females**
Jacob	20,153	1.0013 %	Sophia	21,695	1.1297 %
Mason	19,396	0.9637 %	Isabella	19,745	1.0282 %
William	17,151	0.8522 %	Emma	18,674	0.9724 %
Jayden	16,861	0.8378 %	Olicia	17,169	0.8940 %
Noah	16,719	0.8307 %	Ava	15,383	0.8010 %

A.3. Dataset coverage

An important limitation of the application in Mislove et al. (2011) is mentioned by the authors. Their name-gender dataset can cover different fractions of the male and female populations. Covering different fractions raises a chance of a bias in the outcome of the classifier, as a certain gender is more likely to be recovered than the other.

In order to examine the coverage of the different datasets, the total amount of born babies per gender has to be known. From this information the coverage of the various name-gender datasets can be calculated. To obtain these totals, the top 1000 most popular names of the SSA is used. They provide the top 1000 most popular male and female names together with either an absolute number or a percentage of the total number of births for that specific gender. Combination of both these tables can be used to calculate the total amount of babies per gender in a specific year. A sample of such a combined table is shown in table A.3.

In the work of Li (2012) a similar approach has been performed to calculate the total amount of births per gender as will be done here. He calculates the total amount

by picking the most popular name and divides the number of times it was given by the corresponding percentage of all births. Here a slight modification is made to this approach. The cumulative values of all top 1000 names are used per gender.[1] This minimizes the effect of rounding errors in the dataset of the SSA and can be up to a factor 1000 more precise as effectively the average amount over 1000 measurements is used. The total number of births per gender and year is defined as

$$n_{total}(y, g) = \frac{\sum_{\text{top } 1000} d(y, g, n)}{\sum_{\text{top } 1000} p(y, g, n)}, \tag{A.8}$$

where $p(y, g, n)$ is the percentage the name has been given to a baby of gender g in year y.

A figure showing the coverage of the different name-gender datasets with respect to the U.S. population per year is shown in figure A.2. Here it is seen that a larger variance in female names over the last years results in a smaller coverage of the female population of the U.S. compared to the male population. This deviation shows more clearly in the strict and loose datasets. This is because a larger variation in names goes together with a smaller amount of times a name is given, making the name less accurate and more likely to be excluded from the datasets. Hence, this deviation is seen to lesser extent in the neglect dataset since the unknowns are set to zero.

A different coverage between males and females can give a bias in the results as a certain gender is more likely to be discovered by the algorithm then the other. Here, males are expected to be discovered more often than females using the SSA dataset. A certain bias towards a gender will be checked for in the next subsection.

In addition, now the total amount of births per year is known from (A.8), the power law estimation can be validated used for the loose 95% set can be validated. The total amount of births is compared with the known values from the SSA data set, complemented by the values given by the power law. Here it is found that the power law distribution underestimates the actual amount of births in the extrapolated area by approximately 256.5% of the estimated value, resulting in a 2.7% smaller amount of estimated total births.

While this does net necessarily means that the estimated $n_{missing}$ is wrong, it does imply that the fitted power law does not describe the extrapolated area accurately. The search for a better fit in the extrapolated region has not been performed since the name-gender datasets are validated using "known" real life name-gender combinations.

[1] Take for instance the name Jacob in table A.3. The percentage of total males is 1.003%, which means it can lie anywhere in [1.0025%, 1.0035%). The error it introduces in the total population is given by

$$\frac{20,153}{1.0035\%} - \frac{20,153}{1.0025\%} \approx 201.$$

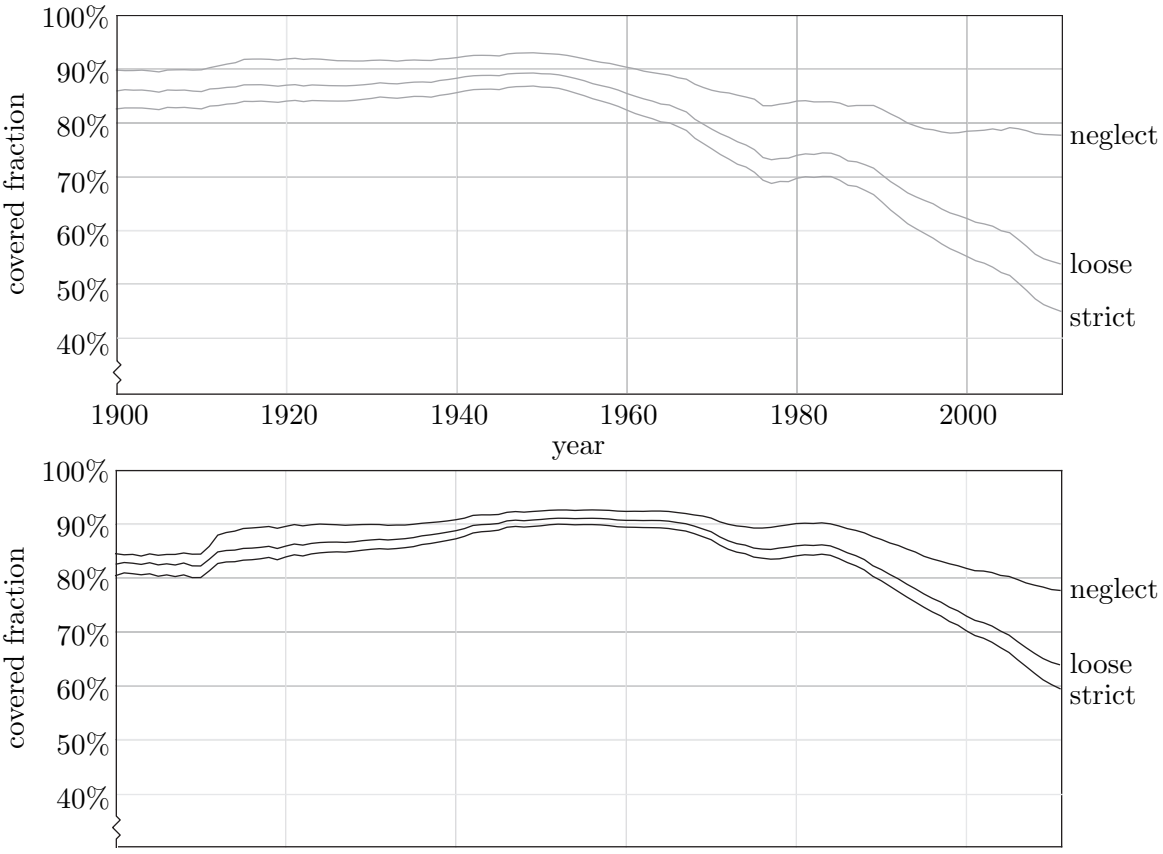


Figure A.2.: Comparison of the gender coverage in percentage over years 1940–2011, where gray is the female coverage (top) and black is the male coverage (bottom). The data is calculated from the SSA data used in this research.