Decision Support Systems 59 (2014) 262–273

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

# External validity of sentiment mining reports: Can current methods identify demographic biases, event biases, and manipulation of reviews?

Fons Wijnhoven *, Oscar Bloemen

*University of Twente, Enschede, Netherlands*

ARTICLE INFO

*Article history:*
Received 5 June 2013
Received in revised form 22 November 2013
Accepted 16 December 2013
Available online 25 December 2013

*Keywords:*
Sentiment mining
Opinion mining
External validity
Demographic bias
Event bias
Product review manipulation
Design proposition validation

ABSTRACT

Many publications in sentiment mining provide new techniques for improved accuracy in extracting features and corresponding sentiments in texts. For the external validity of these sentiment reports, i.e., the applicability of the results to target audiences, it is important to well analyze data of the context of user-generated content and their sample of authors. The literature lacks an analysis of external validity of sentiment mining reports and the sentiment mining field lacks an operationalization of external validity dimensions toward practically useful techniques. From a kernel theory, we identify multiple threats to sentiment mining external validity and study three of them empirically 1) a mismatch in demographics of the reviewers sample, 2) bias due to reviewers' incidental experiences, and 3) manipulation of reviews. The value of external validity threat identifying techniques is next examined in cases from Goodread.com. We conclude that demographic biases can be well detected by current techniques, although we have doubts regarding stylometric techniques for this purpose. We demonstrate the usefulness of event and manipulation bias detection techniques in our cases, but this result needs further replications in more complex and more competitive contexts. Finally, for increasing the decisional usefulness of sentiment mining reports, they should be accompanied by external validity reports and software and service providers in this field should incorporate these in their offerings.



## 1. Introduction

Knowledge of the experiences clients have with a company and their competitors' products and services is crucial, as responding correctly to this information can lead to a competitive advantage [1]. Nowadays, acquiring this knowledge can be assisted by using the ever growing number of sentiments-containing expressions publicly available in (micro-)blogs, review sites, and forums [2]. Manual examination of all these data is a daunting task and automation is desirable.

Solutions for the automated extraction of sentiments come from a subsidiary of machine learning, named opinion or sentiment mining [3,4]. Determining the sentiment of a text is in essence a classification problem with classes positive, negative [2] and neutral [5]. Given an opinionated text as in the book review of Fig. 1, a classifier may determine the polarity of the sentiment by comparing words in the text with words in a lexicon of which the polarity is known. In the case of this book review, words like “great”, “helped”, and “good” indicate a positive review. Analyzing this text with sentiment mining tool Pattern [6] shows that the review is 0.19 positive on a scale from −1 to 1.

The various applications of sentiment mining span a large domain, like movies [4] commercial products and services [3,7,8], product features [9,10] also on a comparative basis [11], and the sentiment toward a political party or topic [2]. In the majority of sentiment mining research, the dominant topic is the classification algorithm. The algorithms are continuously improved to squeeze out the last percentage increase in accuracy [12]. However, if the goal of sentiment mining is harvesting market or public information for decision making, it is of importance to know how the sample corresponds to the target group of which sentiment conclusions are drawn. This problem is known in the field of psychological and sociological research methodology as the external validity of research, which is defined by Shadish et al. ([13], p. 83) as: “… inferences about whether the cause-effect relationship holds over variations in persons, settings, treatments, and outcomes.” In this article, the cause-effect relationship is of type product-sentiment or (political) topic-sentiment and the variations in persons, settings, treatments and outcomes between the target group of which the sentiment is measured compared to the available online sample. If the book of Fig. 1 was written for an audience without a background in mathematics, the book should be evaluated by people from such a population, but from a typical sentiment mining report we do not know if this is the case.

Sentiment mining researchers have only recently started to acknowledge the problem of external validity [14]. Wu et al. [15] acknowledge that there is a problem related to customer group representations and propose a visualization of sentiment mining results including customer groups. In response to the many sentiment mining publications based on Twitter data, Mislove et al. [14] found a Twitter population that was highly deviating from the US demographics. Gayo-Avello [16,17] argues that skewness in demographics contributes to failures

* Corresponding author.
*E-mail addresses:* a.b.j.m.wijnhoven@utwente.nl (F. Wijnhoven), oscarbloemen@gmail.com (O. Bloemen).

0167-9236/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dss.2013.12.005

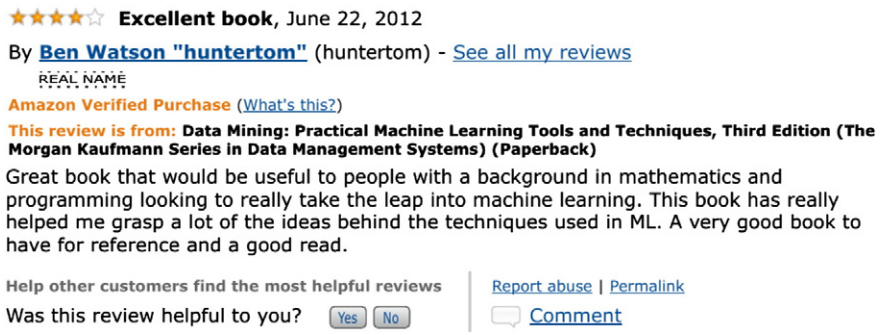


**Fig. 1.** Review over Witten et al. (2011) from Amazon.com, accessed August 16, 2012.

of predicting electoral outcomes from social media content and encourages research toward automatic profiling of social media content authors.

This article attempts to fill this gap on the external validity of sentiment mining by taking into account the context of opinion-expressing authors. We elaborate on context extraction methods, following a product-oriented design theory approach [18], which involves first the detection of its kernel theory, next the identification of its meta-requirements, third the listing of meta-designs for solution artifacts, and finally testing the validity of design propositions. The creation of solution artifacts "… relies on existing kernel theories that are applied, tested, modified, and extended through the experience, creativity, intuition, and problem solving capabilities of the researcher" ([19], p. 76). A product-oriented kernel theory provides ideas for meta-requirements and meta-designs that help to solve classes of problems and create classes of artifacts. Next, design propositions describe effective relations between requirements and designs that can be subject to tests [18,20]. If the design propositions can be corroborated, i.e. sufficient evidence is found that the resulting design does what it is required to do, the bias identifier is expected to be reliable and useful for empirically identifying the size of biases in a sample of sentiment-expressions. If the proposed design propositions cannot be corroborated, no statement about the existence of biases in the sample is possible. Section 2 gives a kernel theory of external validity in social sciences from which three possible sentiment-mining biases are derived. Section 3 gives results of a structured literature review to identify meta-requirements and meta-designs for a sentiment mining biases identifier. Section 4 gives tests of the design and empirical propositions. Finally, Section 5 gives the conclusions and implications of this study.

## 2. Kernel theory: external validity of sentiment mining reports

Shadish et al. ([13], p. 86–90) give five threats to external validity. The first threat (T1) reflects the properties of the sample units, for example the gender and educational level of the people in the sample, and how they relate to the causal relationship. 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. Example is a possible payment for participation in an experiment. The third threat (T3) indicates that findings of a specific study cannot be extrapolated to different outcomes. Shadish et al. [13] here give the example of establishing the effectiveness of a medical treatment, which could be measured in quality of life, 5-year metastasis-free survival, or overall survival. These outcomes may differ and thus cannot be easily generalized to each other. The fourth threat (T4) indicates that observations may be biased by specific settings that do not represent the situations over which one wants to generalize. For example, the test of medical drugs may have different results in developed and developing countries due to different health hazards in both. The fifth threat (T5) is related to the way that causal patterns are identified. The paths that explain causal relationships can be different across various settings. For example the financial crisis in The Netherlands may be reinforced by a too high consumption of mortgages, whereas in Greece it is reinforced by poor government budget control.

Application of these five threats to sentiment mining reveals possible problems with sentiment mining results. From threat T1, the first form of possible bias B1 is due to a mismatch in demographic properties of the sample and target audience, e.g. if the researcher is interested in the public opinion of a specific population, the authors of these opinion expressions must reflect this population. The next problem lies in the motivation of posting a review online. 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.

Threat T2 introduces a problem related to personal experiences of the author, i.e. the sentiment is biased by specific events (B3). Examples include: a review author with a negative sentiment due to certain problems with an old product that would not occur in the new version, or review authors may develop a generally negative attitude due to conditions without any relevance for the product, like the negative evaluations of a movie after several power outages during its presentation.

Threat T3 relates to the type of information that is extracted with the sentiment mining tool. If the interest is an overall sentiment regarding a product, this should be extracted, but if conclusions are drawn about specific features of the product, generalization toward general sentiments may be invalid. This is a kind of analysis error caused by the non-comparability of aspects or features that are mined. Careful selection of the aspects and features in the mining method therefore is a fundamental task for avoiding external validity problems [9–11].

Threat T4 describes the importance of the research setting when generalizing the findings. In sentiment 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 product users to determine sentiments regarding Samsung products is expected to give different results than doing the same on an Android forum. Such platform biases (B4) involve a combination of previously mentioned demographic, manipulation and event biases.

Threat T5 [21] relates to the causal path that links analysis of sentiments to the sentiments of the author (B5), which lies within the applied sentiment mining algorithm. These paths are typically described by features found using a machine learning algorithm. The majority of publications in sentiment mining research concerns with refinements of these algorithms [12].

Table 1 gives an overview of the relations between threats to external validity in social sciences and possible problems in sentiment mining research. For this article, we focus on biases due to demographics, events, and manipulation.

Using the following SCOPUS query ["opinion mining" OR "sentiment analysis" *OR* (*Mining AND* ("social media" *OR* "user generated content" OR reviews OR blog OR forum*)] we found a large set of relevant literature on sentiment mining. The set was made more specific by extending the query with ["external validity" OR generali* OR sample OR noise OR

**Table 1**
Threats to external validity of sentiment mining reports.

| External validity threat | Presence of biases in sentiment mining |
|---|---|
| T1: Interaction of causal relations with units | B1: Demographics bias<br>B2: Manipulations of reviews |
| T2: Interaction of causal relations with treatments | B3: Bias caused by events |
| T3: Interactions of causal relations with outcomes | This is not a bias but an error in the preparation of the analysis. |
| T4: Interaction of causal relations with settings | B4: platform bias. Different platforms involve multiple biases (B1, B2, and B3) and their interactions. Studying these interactions first requires more certainty among B1, B2, and B3, and thus is a follow up study. |
| T5: Context-dependent mediations | B5: algorithm bias. Causal paths creation has been researched extensively already. Consequently not the focus of this study. |

bias]. The literature search revealed the existence of a gap in research. While some papers can be found that discuss the possible existence of bias sources on sentiment mining [12,14,16,17,22–30], the importance of these bias factors is only explicated by Gayo-Avello [17] and Mislove et al. [14]. Matching demographics of the sample and target audience is researched by Meyerson and Tryon [31] and Ross et al. [32] who have compared Internet surveys with their offline counterpart. They both conclude that Internet survey results have a high correlation with the results of traditional surveys. However, Meyerson and Tryon [31] observed skewed demographics in their Internet sample. A similar caveat is present in the study of Ross et al. [32]. Both studies only found agreement with field research after correcting for the skewed demographics of the online sample. Studies regarding social media also have shown that the demographics of users is not parallel to the population. In MySpace, for instance, females dominate over males and younger people are overrepresented [33–35]. Regarding Twitter, Mislove et al. [14] (p. 21) conclude that "Twitter users significantly over-represent the densely population regions of the U.S., are predominantly male, and represent a highly non-uniform sample of the overall race/ethnicity distribution". These are all interesting insights, however, the only show fragments of the external validity problem, and no previous work has described external validity and its dimensions and appearances in the sentiment mining literature in s systematic way.

Dellarocas [36] argues that anonymous user-generated content, combined with the growing influence of online reviews on consumer behavior, gives stakeholders incentives to manipulate online reviews. The commercial importance of positive evaluations is an incentive for biased reviews of own products or products of a competitor [37]. Chen et al. [38] identified a particularly strong impact of product reviews in the pre product release stage, where it highly impacts on investors' decisions to buy into the product. Luo, Zhang and Duan [39] show large impacts of social media communications and ratings on firm equity value. Zhao, Yang, Narayan and Zhao [40] show that fake reviews increase consumer's uncertainty, because the detection of fake reviews and manipulation is less easy. Hu et al. [24] show that manipulation of reviews is a serious problem, and reveal that just above 10% of the books on Amazon.com have manipulated reviews.

Missen et al. [12] conclude that changes of public sentiments on blogs 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. Similar conclusions can be drawn from the research of Das et al. [41], who apply natural language processing techniques to identify locations, the times at, and event that may have led people to a certain sentiment. The idea of explaining sentiments by events is in agreement with the context model of Greenberg [23], which stresses the changing nature of sentiments as participants are continuously subject to events that influence their sentiment temporarily. A sentiment regarding a hotel may for example be positively influenced by exceptionally beautiful weather at the moment of stay which is not representative for the average condition.

## 3. Meta-requirements and meta-design for author context detection

For each of the three external validity biases, a search of literature has been done to find meta-requirements for a bias detection tool. The query for demographic bias articles is "("sample bias" OR "control variables" OR "generali*) AND age AND gender AND education". The query for event bias articles is "(Event* OR experience* OR topic* OR trend*) AND (Internet OR web OR online OR "user generated content" OR twitter) AND ("opinion mining" OR "sentiment analysis")". The query for manipulation articles is "(Manipulate* OR spam) AND (review* OR "user-generated content*) AND (Detect* OR identify*) AND ("opinion mining" OR "sentiment analysis")". The resulting articles are discussed further in this section.

### 3.1. Meta-requirements for a demographic bias detector

The commonly used variables in social sciences expressed in Babbie [42] include gender, age, location, and education. These variables were also found in the papers we found by our SCOPUS query. The list of papers found is here: [12,14–17,26,28,31–35,43–60]. These demographic variables are expected to be sufficient in 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 them a good starting point for future research. Hence, a demographic bias detector must be able to identify gender, age, location of the author at time of posting the review, and educational level of the author.

### 3.2. Meta-design of a demographics bias detector

For extracting age, gender, location, and education, all the papers mentioned in the previous section provide some bias detection techniques for demographics except [16,17]. The given techniques are of three types: writing style analysis —mostly named stylometry [61]-, profile extraction and name classification.

Different types of information are available from many profile pages. Abel et al. [43] and Balduzzi [46] note that on some social networks all demographic variables are available, i.e. the user's age, gender, location, and education. A problem lies in whether this profile information is trustworthy, as the information is self-provided. For example an analysis of Caverlee and Webb [33] of MySpace users showed that unexpected peaks of self-reported age occurred around ages 69 and 100. While the actual age is unknown, their findings suggest that the peaks occurred due to false profile information.

Usernames are deliberately chosen and therefore possess information about the related persons. Furthermore, given names and family names can reflect gender and ethnicity [14]. An example of this is given in Fig. 2, which is an excerpt of comments on a video on YouTube. The numeric value appended to the username of "TheTorri98" is his or her year of birth as one can deduce from the conversation. The inclusion of a game console in username "XclusiveXbox" is likely an indication

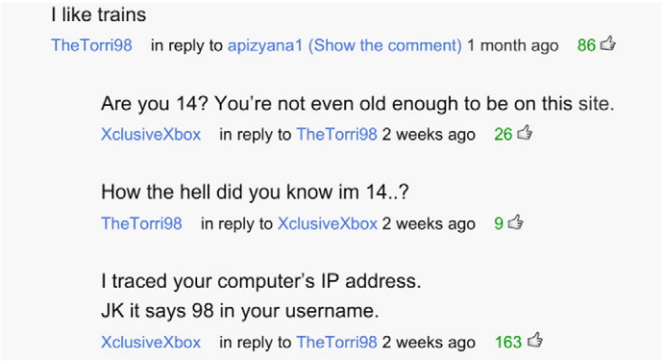


**Fig. 2.** Example of a username with demographic data from, July 2, 2012.

that the user is a gamer and uses Microsoft's X-box console. While gaming is quite domain specific, general demographic features as gender and age can be extracted from name words and numbers like "John" and "98".

The review text itself can be used to collect demographic features of authors. Argamon et al. [45] identify content-based and style-based features for authorship profiling. They call content-based markers problematic because they can be influenced by the writing situation and care is needed in detecting the proper features. The style-based features include function words and typical online blog elements as the usage of hyperlinks and slang, by which Argamon et al. [45] were able to extract gender, age, native language, and neuroticism with 76.1%, 77.7%, 82.3%, and 65.7% accuracy respectively.

### 3.3. Meta-requirements for an event-bias detector

Events can influence the sentiment of a person about a certain topic. Particular events might be of interest, such as a new version of a product or maintenance and revision of a service. Events are proposed to be extractable by natural language processing of review texts and related documents or by using groups of texts to cluster similar topics and discover trends over time. See the list of papers we found by our query for event-bias detection [12,41,62–77].

Regarding natural language processing techniques, Abe et al. [62] describe extraction of events experienced by the author from blogs texts. In their research they acknowledge a topic object of the event, semantic type, and factuality. Semantic type indicates the sentiments related to the event and factuality defines if the event is hypothetical or when the event took place. Saurí and Pustejovsky [73] researched the factuality of events by natural language processing techniques and identified their impact on sentiments.

The identification of large events relies on changes in user-generated content over time [78]. Landmann and Zuell [67] identified events by word frequencies in newspaper articles, similar to Tsolmon [79] who identified large events in Japan by noting changes in word usage frequencies in Twitter. Becker et al. [63] described the identification of large events as clustering problems. A collection of user-generated content is grouped together if the description of events matches enough. Furthermore, events are described by a date and time, as well as location [62,63]. Hence, an event identifier must identify 1) location in terms of longitude, latitude, and accuracy, 2) date and time of the event, 3) reach as an indication for the time span of the event, and 4) sentiment effect as the effect of the event on the sentiment of the author.

### 3.4. Meta-design for an event bias detector

Word features in review texts can be used to identify parts-of speech where an event is indicated [62,72]. For example, in the expression "when my car broke down…", "when" refers to the possible event of a car that broke down. A lexicon consisting of event type words can be used to identify candidate events from the review text itself.

However, identifying candidate events is only the beginning [62,65]. After finding candidate events in the text, the corresponding variables of the meta-data have to be determined, like a specific date of the event. This is sometimes complex because (1) notations of dates differ across cultures -for example the European and USA conventions of day/month/year notations, and (2) relative date and/or time indications, such as "yesterday", require the date and time of posting or last update for finding the absolute date and time of the event. Other meta-data of events may consist of the duration of a higher frequency of posting after a certain data, the location data of the affected users, the average sentiment change after a certain moment, and the reach as the number of reviews affected by the event. With respect to finding relations between events and sentiments, an individual positive or negative experience might not be that interesting in the overall results. However, a significant amount, in the sense that it results in a non-negligible change in the public sentiment, is of interest.

By clustering, similar reviews can be grouped based on certain features. If the interest is in many reviews influenced by the same event, it is expected that descriptive words related to that event will have a larger usage frequency compared to unaffected reviews [67]. Examples of such descriptive words may be "President Obama's White House speech", "the G20 Summit", "The Drive movie release".

### 3.5. Meta-requirements for a manipulation detector

The articles found for manipulation are given here [24,25,36,80–87]. The threat of manipulation of reviews was investigated by Dellarocas [36], who describes an easy way to manipulate by a template text that can be filled with details and posted everywhere. By using such a template, the resulting reviews are near-duplicates [80]. Jindal and Liu [25] use near-duplicates to create a dataset of manipulated and not-manipulated reviews. The presence of near-duplicates also reduces the natural randomness expected in online reviews, enabling the detection of manipulated threads by examining the overall statistical properties [24].

A manipulator might show a strong bias toward the product or brand for which it is manipulating [25,81,84,85]. Thereby even downplaying competitive products [25]. Hence, a manipulation identifier must be able to identify the following biases 1) near duplicates, 2) user product or brand bias, i.e. the number of reviews the author has posted on the same site over the same item, 3) polarity deviation, which is the difference of the polarity of this review with respect to the overall polarity of the product, and 4) impact, which is a measure of the impact of the review on the "average sentiment" at time of posting. Different researchers have found that manipulative reviews are posted close to the launch of a product [25,84,86]. Motivations for manipulators are (1) early posts can influence the subsequent review posts and (2) an early manipulated review has a relatively large impact [84]. Jindal and Liu [25] show that a single review for a product is most likely manipulated, followed by the first and second review in case of more reviews.

### 3.6. Meta-design of a manipulation detector

From the literature we identify that manipulation indicators work as follows:

- Near-duplicates: The detection of near duplicates is done using a shingle method [88], which splits a document into different n-gram features—fingerprints of the document. The document is represented by a list of N fingerprints, on which it compares to the fingerprints of other documents. This only makes a first selection of candidates that have a higher chance of being near-duplicates. So you still have to perform some checks on the candidates to be sure, which we do

by calculating the normalized "distance" between the feature sets of two texts. Normalizing this distance gives then a measure for the similarity.

- Metrics for product bias. A product bias is the number of reviews from the user in question divided by the total number of reviews for that product. A bias for a specific brand is the number of reviews for a brand multiplied by the average user sentiment and divided by the total number of reviews for that user. This metric makes it possible to identify users that post many reviews for a particular brand together with an extreme sentiment score (positive or negative). Comparison of this score for different brands can then reveal a large score for the brand the user might be manipulating for, and a low score for competitors.
- Polarity deviation. The polarity deviation is calculated by expressing the absolute difference between the sentiment of the review under investigation and the average sentiment score in terms of the standard deviation of the sentiment scores for the given item.
- Impact. The impact factor is defined as one over the review number (first review being 1, second 2, etc.). The more reviews that have been published, the less the impact of a biased review on the total extracted sentiment of the sample.

## 4. Testing of propositions

### 4.1. Propositions

On basis of the literature review, we are able to state a few relevant design propositions and empirical propositions regarding demographic, event, and manipulation biases of sentiment analysis reports. These are listed in Table 2. For each of these propositions we aim at collecting evidence for identifying the three types of biases using the previously described methods. Regarding the design propositions, the evidence is related to testing the reliability of tools for bias detection. Regarding the empirical proposition, the evidence aims at finding demographic, event and manipulation biases that impact the sentiment discovered in the sample of reviews by applying a tool or combination of bias detection tools.

### 4.2. Data collection

We test the empirical and design propositions using data from Goodreads. Goodreads is a popular online book review website launched in 2007. Its population of reviewers now surpasses 14 million reviewers and they have collectively taken more than 470 million reviews. The availability of meta-data in user profiles makes Goodreads useful for testing the influence of bias (empirical propositions) and testing the quality of proposed extraction algorithms (design propositions). The Application Programming Interface (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. By this method, however, only about 75% of all the reviews are accessible.

To investigate profile bias, a case is sought with a large number of reviews together with rich profile information. For this case, the set of reviews for "The Da Vinci Code" by Dan Brown is chosen, containing 23,526 reviews from 23,505 unique user accounts. Of these users, 19,707 have indicated their gender (83.8%), 8741 have specified their age (37.2%), and 20,664 (87.8%) of the users provided location data on their profile. The large sample and known genders and ages make it possible to train the machine learning algorithm for writing style analysis. Furthermore, the rich profile dataset gives opportunities to test determining gender from given names. Finally, we used the GeoNames database to locate reviewers and to produce a distribution of reviewers and sentiments on a map. A comparison of the sentiment results for different sample groups (male, female, young, old, etc.) from the profile data and the profile extraction methods may identify demographic biases.

To determine the influence of events on public sentiments, a case is needed where it is known that an event happened and that this event had, most likely, an impact on the case. Here, the book "Drive" by James Sallis is chosen. The book was published in 2006 while the movie came out in 2011. This time span gives time for reviews to appear before the event happened, and therefore are known not to be influenced by the movie. The movie being premiered in 2011 still allows for review authors to be influenced by the movie. Furthermore, as with the Da Vinci case, here too the review author profiles are rich in information. There are 265 reviews, all from unique user accounts, 223 which have their gender specified (84.2%) and 107 who have shared their age (40.4%). Here an over-representation in gender is less profound, males are the larger group with 143 user accounts (54.0%) compared to the 80 known females (30.2%). Extrapolation of the gender distribution to the unknown genders yields 64.1% males versus 35.9% females. With respect to event bias, our interest is in (1) detection of the movie launch, (2) examining the number of reviews that are

**Table 2**
Propositions.

| Design propositions | Verdict |
|---|---|
| Profile data can be used to find age | + |
| Profile data can be used to find gender | + |
| Profile data can be used to find location | + |
| Profile data can be used to find education | No data available |
| (User)name analysis can be used to find age | + |
| (User)name analysis can be used to find gender | + |
| Stylometry analysis can be used to find age | − |
| Stylometry analysis can be used to find gender | − |
| Post-frequencies can identify an event | + |
| Data and time of posting can be used to find influenced reviews | + |
| Word-frequencies can identify an event | + |
| Word features can be used to find influenced reviews | + |
| Sentiment mining can find sentiment characteristics related to events | + |
| (Near-)duplicates identify manipulated reviews | + |
| User-brand bias can identify manipulated reviews | + |
| **Empirical propositions** | **Verdict** |
| Demographic bias make sentiment reports invalid for the target group. | + |
| Some events make sentiment reports non-representative for the product or service. | + |
| Some events change the sample of sentiment-expressing authors. | + |
| Sentiment-mining datasets contain manipulated reviews that modify sentiment reports toward a specific sentiment polarity. | − |

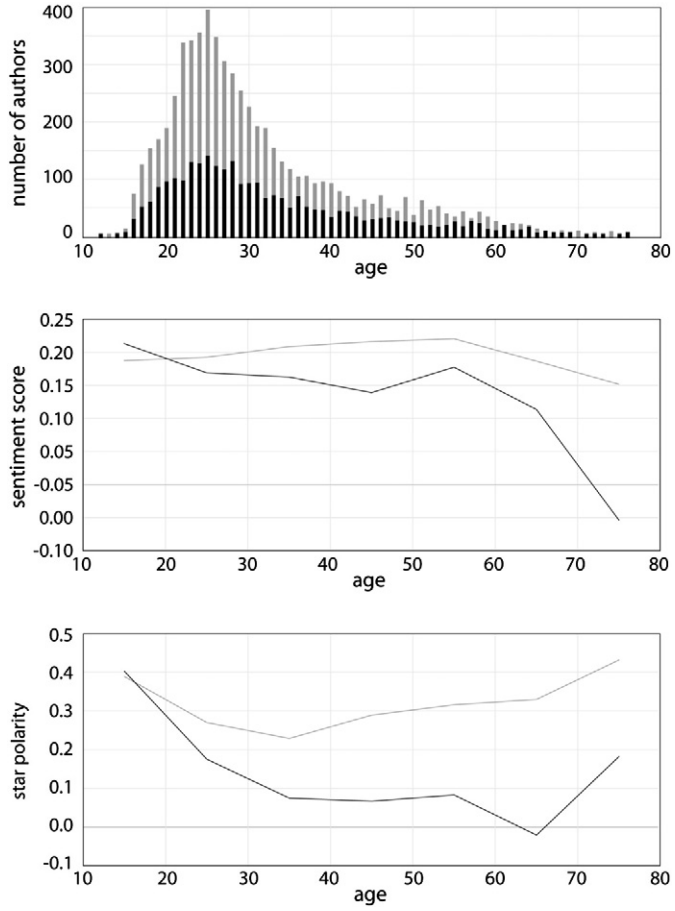


**Fig. 3.** Self-provided age, gender and sentiment of the Da Vinci case. Gray and black lines present females and males respectively. Second graph gives sentiment scores based on author reviews, averaged per ten year. The lowest graph gives normalized star scores by authors.

affected by the movie, (3) possible changes in public sentiment regarding the event, and (4) changes in population of the reviews due to the event.

To investigate the possibility of manipulation, a group of competing products has to be found. We chose a group of books related to the financial crisis of 2008. A list of the 20 "most popular" books on this topic is obtained from Amazon.com. Next, we collected the reviews of these books from Goodreads. Here the interest is in finding (1) (near-) duplicate reviews and (2) promoting and dis-crediting reviews from a single author on different books.

### 4.3. Testing demographics bias

The self-provided age in combination with gender for the Da Vinci Code reviews is shown in Fig. 3. The age distribution of both genders is roughly similar and covers ages 17 until 77. The median age is at 25 for both genders.

The sentiments of both genders toward 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. To check whether this sentiment distribution is an artifact of the sentiment classifier, rather than true sentiment, the self-provided star ratings are normalized and compared with the scores from the sentiment classifier. One can see similar characteristics in the middle graph created by the sentiment classifier as in the lowest graph created by the self-reported stars in Fig. 3. Here too, the increased separation of sentiments between genders as the subsample is older is present, and peak around 50–60. Overall, the ratings are higher when looking at the star-ratings compared to sentiment classification.

In the "Da Vinci Code", gender can be well identified by first name in most cases (see Table 8). To find names that are relatively unique for a gender (at least 95% in favor of a gender), we used name information of the US Social Security Administration 2012 National data. They provide lists of how many babies with a specific gender were given what first name in a specific year. A first name does not always uniquely identify one gender and the SSA only provides name and gender if the name was given at least 5 times to babies of that specific gender in that specific year. Thus, if in 2010 the name James was given 4 times to females, we don't know that and thus have "unknowns". The frequency of the name James for females for that year thus can be between 0 and 4. Using this information a 95% accurate name-gender set can be calculated. Even when we would leave out these unknowns, and assume it to be zero, the precision of gender estimation is still high (see Table 3).

The distribution of males versus females differs between the users that provide their gender on their profile and the users that do not provide their gender (where the gender is found by analyzing the name). 34.2% of the unknown user genders could be recovered by the name algorithm. Out of these, 59.3% were found to be female (compared to 73.6% of the users that did indicate their gender). We also calculated the sentiment scores for this sample and found an average .17 for the known males and average .18 score for the males discovered by the algorithm. For the known females the average sentiment score was .21 and for the females discovered by the algorithm the average sentiment was .23.

Likewise, we tried stylometric analysis to detect demographic information of a user. A multi-Nominal Naïve Bayes classifier is trained by learning textual features from a training set of texts with known gender and age profiles. The learning of features and corresponding probabilities is implemented here by counting the occurrences of individual words per class. Of these possible features the top N features with largest differences in occurrences between the classes is chosen. The top ten extracted features from the different training set are shown in Table 4.

This top features set is roughly similar across the different training sets and show similarities with literature [45]. The age bin "young" is not shown as the Da Vinci Code case does not provide authors within this category. Validation of the classifier is done by comparing the output of the classifier with the self-provided gender and age. The results of this validation are shown in Table 4 for all training methods on the blog dataset. While the precision and recall values are good for overall gender and age classification, the precision and recall values for the individual classes introduce problems. 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

**Table 3**
Precision and recall of gender estimation by first name information in the Da Vinci dataset.

| | 95% certain gender identification | | Neglecting unknowns from the SSA data | |
|---|---|---|---|---|
| | Precision | Recall | Precision | Recall |
| Female | .991 | .630 | .987 | .803 |
| Male | .997 | .643 | .991 | .810 |

**Table 4**
Precision and recall for gender and age using the Multinomial Naïve Bayes classifier as %.

| | Top ten feature words for gender and age | Numbers of reviewers | Precision (P) & recall (R) |
|---|---|---|---|
| Gender | | | P 69.4; R 99.3 |
| Female | I, it, was, book, this, read, loved, me, and, movie | 841 | P 85.8; R 98.6 |
| Male | Of, is, the, a, that, Brown, are, in, Dan, novel | 2593 | P28.8; R 94.3 |
| Unknown | | | 314 |
| Age classification | | | P 72.2; R 99.4 |
| Younger age | Awesome, amazing, Dan, Brown, | 2392 | P 25.0; R 92.6 |
| Older age | very, is, well, overrated, superb, novel the, read, loved, this, fun, I, was, and, page, turner | 11,632 | P 88.6; R 99.3 |
| Unknown | | 740 | |

and males with low precision, both having high recall, then the classifier produces a large overestimation of males. Consequently, this type of stylometry is in our case therefore not reliable for identifying gender and age.

Finally, we used the GeoNames database and were able to relate the reviewers and their sentiments with geographic places. The output of this can be presented in a map as we do in Fig. 4, which clearly shows the global bias of sentiments of the book.

This result is of course according to expectations, because obviously more people read English books, but nevertheless the tools shows that it is feasible and reliable to visualize geographic biases by location names.

### 4.4. Testing manipulation

Since there are no available review datasets that store information on manipulation, a collection of reviews from related books is investigated. This collection of reviews allows for testing the different variables that can make a review or user account suspicious.

Different researchers have used near-duplicate reviews as an indication for manipulated reviews. In the shingle method a document is split up into a feature set. This feature set typically consists of n-grams from the document. Here, the features are 1-grams. The unique features are each given their own dimension and they are given a score of the normalized usage frequency (0 being not used, 1 being the only feature used in the document). This makes the review document a vector of features. The distance between two review-vectors can then be used to calculate a candidate near-duplicate by the maximum distance and Euclidean distance, and normalize it. This results in a score of 1 for documents that share all features, with the same usage occurrences. The method is applied first on the collection of reviews for each book separately. Reviews are marked as near-duplicates candidates when their near-duplicate score is at least .9 (90%). The results are shown in absolute numbers as well as percentages of total reviews per book in Table 5. 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, i.e. “The Big Short” by Lewis, there were 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.”. The other near duplicate texts, also for the ones found for “13 Bankers” by Johnson and Kwak and “House of Cards” by Cohan, are all longer reviews.

As far as the impact of a review score on the average, most are negligible and have 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 a regular user, the second is from a blog on book reviews. The others of these higher impact reviews also appear to originate from other blogs on book reviews.

Considering the set of related books available to test manipulation, a manipulator could post similar reviews amongst different books. The purpose of this can be downplaying sentiment about books while, possibly, advocating the one that the person is manipulating for. A method to detect this is comparing all the reviews available.

Application of the near-duplicate algorithm identified 29 near-duplicates, which corresponds to 0.7% percent of the 3981 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.

Another method to observe manipulation is by examining posting behavior and identifying a brand bias. To start, an overview is created

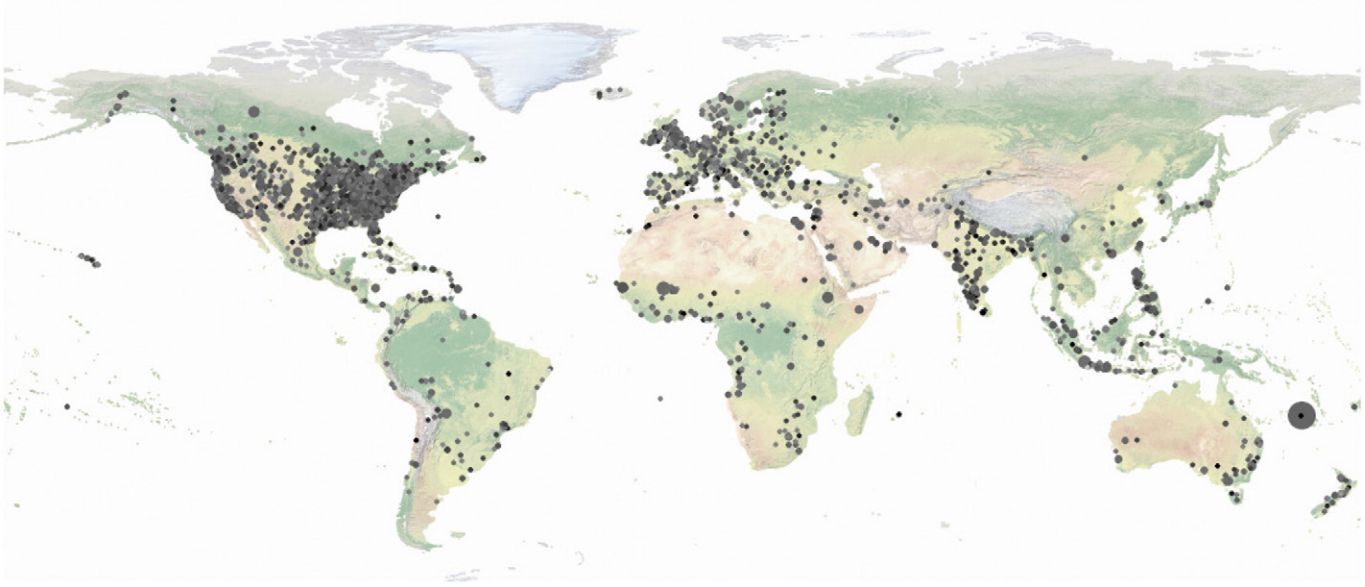

**Fig. 4.** Location and sentiment with black indicating negative sentiment, white indicating positive sentiment, size of circles indicating the number of reviewers per location.

**Table 5**
Number of near-duplicate reviews per book in absolute numbers and % of total.

| Book rank | # of reviews | % Near duplicates | Book rank | # of reviews | % Near duplicates |
|---|---|---|---|---|---|
| 1 | 9 | .4 | 11 | 0 | 0 |
| 2 | 0 | .0 | 12 | 0 | 0 |
| 3 | 0 | .0 | 13 | 0 | 0 |
| 4 | 0 | .0 | 14 | 0 | 0 |
| 5 | 2 | 1.8 | 15 | 0 | 0 |
| 6 | 0 | .0 | 16 | 0 | 0 |
| 7 | 2 | 1.4 | 17 | 0 | 0 |
| 8 | 2 | 25.0 | 18 | 0 | 0 |
| 9 | 0 | .0 | 19 | 0 | 0 |
| 10 | 0 | .0 | 20 | 0 | 0 |

of how users have posted multiple reviews for different books. The results are present in Table 6. The table shows how multiple postings from an account related to the books. Per connection, the percentage of the total amount of reviews corresponding to the book of the column is shown. 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 another 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.

For detecting if the users that post reviews for different books advocate one book in particular, the polarity deviation of the reviews compared to the average for the book is calculated. This deviation is monitored for all reviews of such an user, looking for large differences between the books. Manual inspection of 892 posts from users that posted multiple reviews in the set did not reveal suspicious behavior. From these reviews, 49 referred to one of the other books in the collection, sometimes recommending one (19 times) or stating that the book would be a good supplement (22 times). By far, "The Big Short" takes part in these comparisons (32 times), followed by "Too Big to Fail" (22 times). Furthermore, "The Big Short" is often compared with previous works of Lewis.

As Goodreads.com also provides a user-provided star rating, we also checked if a specific bias toward a book could indicate manipulation. Most of the time, a reviewer likes a book but thinks the other is slightly better, but no product bias by polarity deviation existed in the dataset.

### 4.5. Testing event bias

The Drive case is selected to find if the film had an impact on the book reviews. About one and a half month after the mid 2011 USA release of the movie Drive, the postings show a dramatic increase. In the years before about six reviews per year were given, while after the movie release that same number is easily made in a month time. A second burst in the number of reviews is seen starting from January 2012. A possible explanation for this can be reading during the Christmas holidays. As proposed by Landmann and Zuell [67], events can be extracted by monitoring the frequencies of words over time. The idea is that an event influencing a significant amount of review authors can introduce a sudden rise in the usage of the word or word group that is relevant to the event. Hence, deviations in word usage frequencies have to be investigated for automatic extraction of events. A method to analyze the words that undergo a large change in usage frequency is to calculate the derivative of the usage frequency with respect to time, $\frac{\delta}{\delta t} f_{word}$, where $f_{word}$ is the word usage and $t$ time. The word-frequency data can be smoothed using a Hanning-window of 81 width, hence, searching the largest slope will result in words that undergo a large and sudden change. The words with largest deviations in usage frequency over time are given in Table 7.

From the similarities in the use frequencies of 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 contrary, shows an inverse of the other usage frequencies, perhaps this word is not used by moviegoers. Fig. 5 gives the frequency of feature words before and after the events.

The next question is whether the event of a movie has an impact on the sentiment and whether the event attracted a different audience to the book. To investigate this, different methods that separate the sample into a group influenced by the event and a group not influenced by the event. The first method is selecting the influenced group date of posting by using one of the words "movie", "Gosling" (actor in the movie), and "different". The second method divides the sample into two groups by

**Table 6**
Overview of the postings from the same account for different books.

| | 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 | L0 | 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 |

Note: The numbers indicate how many users have written a review for book X and also wrote a review for book Y as percentage of the total amount of reviews for the book with the column's rank.

**Table 7**
Maximum slope of the word feature, scaled to the word features "Movie" for the Drive case.

| Feature | Movie | Book | Noir | Read | Different | Character | Characters | American | Short | Time |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1.00 | .52 | .59 | .46 | .48 | .37 | .35 | .33 | .32 | .31 |

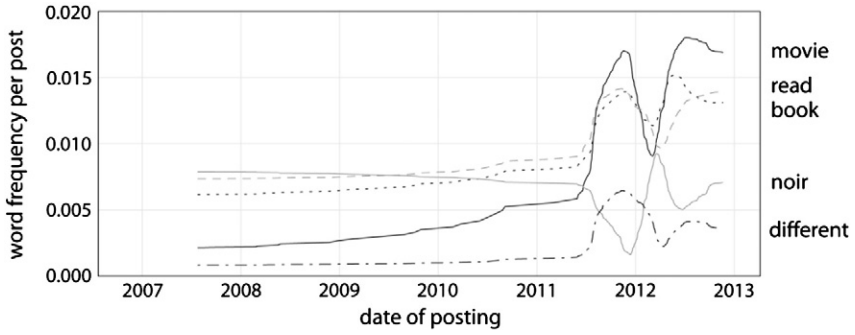


**Fig. 5.** Posting peaks.

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. A comparison of both groups created by the different methods is given in Table 8.

While the small group size makes drawing conclusions difficult, it is interesting to see that the same impacts 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 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. The correspondence between the different selection methods indicates that the event can indeed influence on the sentiment.

## 5. Conclusions

Because sentiment mining harvests public information for decision making, it is important to know how the sample of sentiment expressing people corresponds with the target group for whom the related decisions will be made. This correspondence problem is named the problem of external validity of research. Just recently, authors have started to pick up the challenge of analyzing sentiment mining report's external validity. We contribute to this new field by introducing a conceptualization of external validity, its dimensions and using this as a means to classify the diverse fragmented attempts for tool development. Next, we created a survey of existing methods for establishing possible demographic, event and manipulation external validity biases of sentiment mining reports, and we tested their usefulness by book reviews in Goodreads.com.

In the literature we found two methods for the extraction of gender and age biases. The first method estimates gender by the reviewer's first name. The second method estimates gender and age from the writing style. Using data from reviews of The Da Vinci Code, we estimated gender from the user's first name, which showed highly accurate results and was able to recover large parts of the test set. While reasonable accuracies were obtained by the application of stylometry, the method produced a bias, i.e., it overestimated the number of males and older age people in the author population. This bias makes the stylometric analysis techniques we used not yet usable for recovering missing genders and ages, and alternative stylometric techniques are needed to be able to use stylometry reliably. For the identification of location, we used the GeoNames database for converting place names to geographic coordinates and for presenting the number of reviews and sentiments on these coordinates. Using the self-provided information of the review authors it was shown that as the reviewers get older, the differences between males and females become more prominent. Around ages 40–59 the women are, on average, roughly 30% more positive than the men. The Geographic distribution of reviews and sentiments over the globe were highly biased, making sentiment mining results from GoodRead data less representative for many areas in the world.

Two methods were used to monitor the effects of events on the sample. Book reviews for the book "Drive" by James Sallis were chosen. By observing changes in the posting frequency, we could successfully identify a large peak shortly after the "Drive" movie's premiere. Furthermore, using a clustering approach by analyzing word usage frequencies, the "movie-event" was successfully discovered. The event appeared to

**Table 8**
Comparison of book reviews influenced by the movie (inf.) and authors not influenced by the movie (not-inf.) using two group separation methods.

| | Groups separation method | | | |
|---|---|---|---|---|
| | Word | | Time | |
| | Influenced | Not-influenced | Influenced | Not-influenced |
| Group size | 151 | 113 | 45 | 23 |
| Average sentiment | 0.20 | 0.13 | 0.15 | 0.12 |
| Male % (males) | 59 (75) | 71 (67) | 62 (26) | 90 (19) |
| Average age (known values) | 34 (64) | 39 (43) | 33 (22) | 55 (9) |

modify the sentiment, as well as population that posted reviews. The reviewers influenced by the movie-event appear younger, are more often female, and have a more positive review text.

To detect manipulation, the reviews of a collection of books covering the financial crisis of 2008 were used. By analyzing near-duplicates, suspicious reviews were identified. Some of the shorter text, most notably common sayings as "Loved it!", appeared as duplicates. Whether these texts are posted with the intention to manipulate is doubtful, as the impact of these reviews on the average rating in our sample is negligible (< 1%). From this research it appeared that duplicate reviews, while suspicious, do not necessarily mean manipulation. Furthermore, users who posted reviews for different books in the financial crisis dataset were investigated. Their reviews were checked for a bias toward a specific book, but no polarity deviations were found. Thus although candidates of manipulated reviews can be detected by existing techniques, they do not necessarily imply manipulative behavior and in fact we had no evidence of manipulation in the dataset.

Before interpreting these findings, we need to mention a possible representation bias in our own data. The Goodreads API only allowed for collecting approximately 75% of all the reviews available per book. In the API documentation GoodRead says that the most popular reviews are returned. Their algorithm for determining the popularity of a review is unrecoverable for us but could introduce a bias in the presented results. This means that our study needs replication from a corpus that lacks these unknowns. Future research may also extend this work by enriching missing profile data by searching for multiple profiles related by a similar username. This would further elaborate on the work of Abel et al. [43] and Perito et al. [89]. Furthermore, sources inside a corporation, like transaction history and customer relationship data [90], could be accessed to gain more information or improve accuracy.

Following the summarized results in the verdicts of Table 2, we have the following limitations and needs for further research. We did not study educational level indicators, because they were not present in the GoodRead database. Following insights from sociological research and marketing, educational level is an important influencer of sentiments and thus the inclusion of educational indicators is important for the evaluation of the external validity of sentiment mining reports. If profile data or self-provided data on educational level are not available, proxies could be used like the link between educational level and location as is present from national census data. Writing style analysis may be also a useful indicator of educational level, indicating an even more urgency for research on stylometric analysis methods, not only for identifying age and gender but especially for identifying educational levels. Finally, we did not find any empirical evidence of manipulation in the GoodRead database. Given the likelihood of manipulation on other platforms where more competitive products are reviewed and less control exists on who is reviewing what and how, this study should be replicated.

One may criticize our review of event bias as too simplistic. We agree, although book reviews are real product reviews. Because the introduction of external validity issues in sentiment mining is rather new and under researched, we preferred to keep the application part in this article simple. Elaborating on the complexities involved in other cases, let us take hospital service reviews as an example. If a hospital receives much lower sentiments, finding the reasons for this may be a big puzzle varying from events such as the announcement of new treatments, some statements of dissatisfaction on a microblog, the appointment of a new doctor or CEO, and reviews published by the national health authority and insurance firms. A more intensive analysis of review texts can deliver the possible causal relation to any of such events, but often multiple events work together [23]. Additionally, a lot of reviews are needed to do such an exploration effectively, which merits a different study of its own.

For future research we also 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. Furthermore, different websites should be used to search for differences in sentiment with respect to the website setting, which can show interesting interactions of demographic, event and manipulation biases, i.e., platform biases. Some platforms may better control for them or make biases more transparent in reports.

We believe that the value of this research for practice may be multiple. First, although sentiment-mining can deliver important insights in the word-of-mouth regarding products and services [2], the impact and importance of these word-of-mouth insights need to be reliable and should not point decision makers in the wrong directions. This therefore requires that sentiment-mining reports have to be supported by additional insights in their external validity. Secondly, we did not find any software or service provider in the brand monitoring business that delivers external validity reports of sentiment mining. This makes sentiment mining not well interpretable and the software and service providers who can deliver these external validity reports may be much in advantage and have more value to deliver than those that do not.

## References


[1] J.C. Narver, S.F. Slater, The effect of a market orientation on business profitability, The Journal of Marketing 45 (4) (1990) 20–36.
[2] B. Pang, L. Lee, Opinion mining and sentiment analysis, Foundations and Trends in Information Retrieval 2 (1–2) (2008) 1–135.
[3] K. Dave, S. Lawrence, D.M. Pennock, Mining the peanut gallery: opinion extraction and semantic classification of product reviews, Proceedings of WWW, 2003, pp. 519–528.
[4] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the Conference on Empirical Methods in Natural, 2002, pp. 79–86.
[5] M. Koppel, J. Schler, The importance of neutral examples for learning sentiment, Computational Intelligence 22 (2006) 100–109.
[6] T. De Smedt, W. Daelemans, Pattern for python, The Journal of Machine Learning Research 98888 (2012) 2063–2067.
[7] P.D. Turney, Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the 40th Annual Meeting of the Association for, Computational Linguistics (ACL), 2002, pp. 417–424.
[8] M. Sokolova, G. Lapalme, Learning opinions in user-generated web content, Natural Language Engineering 17 (2011) 541–567.
[9] X. Ding, B. Liu, P.S. Yu, A Holistic lexicon-based approach to opinion mining, Proceedings of the International Conference on Web Search and Web Data Mining, 2008, pp. 231–239.
[10] C. Brun, Detecting opinions using deep syntactic analysis, Recent Advances in Natural Langiage Processing, Hissar, Bulgaria, 2011.
[11] K. Xu, et al., Mining comparative opinions from customer reviews for competitive intelligence, Decision Support Systems 50 (2011) 743–754.
[12] M.M.S. Missen, M. Boughanem, G. Cabanac, Opinion detection in blogs: what is still missing? 2010 International Conference on Advances in Social Networks Analysis and Mining, IEEE, Odense, Denmark, 2010, pp. 270–275.
[13] W.R. Shadish, T.D. Cook, D.T. Campbell, Experimental and Quasi-Experimental Designs, Houghton Mifflin Company, New York, 2002.
[14] A. Mislove, et al., Understanding the demographics of Twitter users, in: N. Nicolov, J. Shanaha (Eds.), Fifth International AAAI Conference on Weblogs and Social Media, AAAI Digital Library, Barcelona, 2011, pp. 17–21.
[15] Y. Wu, et al., OpinionSeer: interactive visualization of hotel customer feedback, IEEE Transactions on Visualization and Computer Graphics 16 (2010) 1109–1118.
[16] D. Gayo-Avello, P.T. Metaxas, E. Mustafaraj, Limits of electoral predictions using Twitter, in: N. Nicolov, J. Shanaha (Eds.), Fifth International AAAI Conference on Weblogs and Social Media, AAAI Digital Library, Barcelona, 2011, pp. 490–493.
[17] D. Gayo-Avello, A meta-analysis of state-of-the-art electoral prediction from Twitter data, Arxiv preprint arXiv:1206.58512012.
[18] J.G. Walls, G.R. Widmeyer, O.A. El Sawy, Building an information system design theory for vigilant EIS, Information Systems Research 3 (1) (1992) 36–59.
[19] A.R. Hevner, et al., Design science in information systems research, MIS Quarterly 28 (1) (2004) 75–105.
[20] M.L. Markus, A. Majchrzak, L. Gasser, A design theory for systems that support emergent knowledge processes, MIS Quarterly 26 (2002) 179–212.
[21] W.R. Shadish, T.D. Cook, D.T. Campbell, Experimental and Quasi-Experimental Designs, 2002.
[22] A. Das, S. Bandyopadhyay, Towards the Global SentiWordNet, Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, 2011, pp. 799–808.
[23] S. Greenberg, Context as a dynamic construct, Human-Computer Interaction 16 (2001) 257–268.

[24] N. Hu, et al., Manipulation of online reviews: an analysis of ratings, readability, and sentiments, Decision Support Systems 52 (2012) 674–684.
[25] N. Jindal, B. Liu, Opinion spam and analysis, Proceedings of the International Conference on Web Search and Web Data Mining, ACM, New York, 2008, pp. 219–229, (New York).
[26] J. Oberlander, S. Nowson, Whose thumb is it anyway? Classifying author personality from weblog text, Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, Association for computer linguistics, Stroudsburg, PA, 2006, pp. 627–634, (Sydney, Australia).
[27] B. Stone, M. Richtel, The hand that controls the sock puppet could get slapped, The New York Times (2007).
[28] M. Thelwall, D. Wilkinson, S. Uppal, Data mining emotion in social network communication: gender differences in MySpace, Journal of the American Society for Information Science and Technology 61 (2009) 190–199.
[29] J. van Dijck, Users like you? Theorizing agency in user-generated content, Media, Culture & Society 31 (2009) 41–58.
[30] Q. Ye, Z. Zhang, R. Law, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, Expert Systems with Applications 36 (2009) 6527–6535.
[31] P. Meyerson, W.W. Tryon, Validating internet research: a test of the psychometric equivalence of internet and in-person samples, Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc. 35 (2003) 614–620.
[32] M.W. Ross, et al., Biases in internet sexual health samples: comparison of an internet sexuality survey and a national sexual health survey in Sweden, Social Science & Medicine (1982) 61 (2005) 245–252.
[33] J. Caverlee, S. Webb, A large-scale study of MySpace: observations and implications for online social networks, Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries, 2008, pp. 104–114, (Pitsburgh, PA) Association for the advancement of artificial Intellgence. Paper is accessible via URL www.aaai.org/Papers/ICWSM/2008/ICWSM08-012.pdf.
[34] M. Thelwall, Social networks, gender, and friending: an analysis of MySpace member profiles, Journal of the American Society for Information Science and Technology 59 (2008) 1321–1330.
[35] U. Pfeil, R. Arjan, P. Zaphiris, Age differences in online social networking — a study of user profiles and the social capital divide among teenagers and older users in MySpace, Computers in Human Behavior 25 (2009) 643–654.
[36] C. Dellarocas, Strategic manipulation of internet opinion forums: implications for consumers and firms, Management Science 52 (2006) 1577–1593.
[37] D. Houser, J. Wooders, Reputation in auctions: theory, and evidence from eBay, Journal of Economics & Management Strategy 15 (2) (2006) 353–369.
[38] Y. Chen, Y. Liu, J. Zhang, When do third-party product reviews affect firm value and what can firms do? The case of media critics and professional movie reviews, Journal of Marketing 76 (2) (2012) 116–134.
[39] X. Luo, J. Zhang, W. Duan, Social media and firm equity value, Information Systems Research 24 (1) (2013) 146–163.
[40] Y. Zhao, et al., Modeling consumer learning from online product reviews, Marketing Science 32 (1) (2013) 153–169.
[41] A. Das, S. Bandyopadhyay, B. Gambäck, Sentiment analysis: what is the end user's requirement? Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, ACM, New York, 2012, (Craiova, Romania).
[42] E.R. Babbie, The Practice of Social Research, 11th ed. Cengage Learning, Belmont CA, 2007.
[43] F. Abel, et al., Interweaving public user profiles on the web, in: N. Henze, E. Herder, D. Krause (Eds.), User Modeling, Adaptation, and Personalization, 18th International Conference, 2010, pp. 16–27.
[44] S. Argamon, et al., Gender, genre, and writing style in formal written texts, Text 23 (2003) 321–346.
[45] S. Argamon, et al., Automatically profiling the author of an anonymous text, Communications of the ACM 52 (2009) 119–123.
[46] M. Balduzzi, et al., Abusing social networks for automated user profiling, in: Somesh Jha, Robin Sommer, Christian Kreibich (Eds.), Proceedings of the 13th International Conference on Recent Advances in Intrusion Detection, Lecture Notes in Computer Science, Vol. 6307, Springer, Ottawa, Canada, 2010, pp. 422–441.
[47] F. Can, J.M. Patton, Change of writing style with time, Computers and the Humanities 38 (2004) 61–82.
[48] R. Chandramouli, K.P. Subbalakshmi, Gender identification from E-mails, 2009 IEEE Symposium on Computational Intelligence and Data Mining, 2009, pp. 154–158.
[49] J.R. Buck, S.J. Cheng, Instructions and feedback effects on speed and accuracy with different learning curve models, IIE Transactions, 1993, pp. 34–37.
[50] N. Cheng, R. Chandramouli, K.P. Subbalakshmi, Author gender identification from text, Digital Investigation 8 (2011) 78–88.
[51] M. Dahllof, Automatic prediction of gender, political affiliation, and age in Swedish politicians from the wording of their speeches—a comparative study of classifiability, Literary and Linguistic Computing 27 (2012) 139–153.
[52] D. Estival, et al., Author Profiling for English and Arabic Emails, 2008. 1–22.
[53] K. Filippova, User demographics and language in an implicit social network, Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning, Association for computer linguistics, Stroudsburg, PA, 2012, pp. 1478–1488, (Jeju, Korea).
[54] X. Geng, Z.-H. Zhou, K. Smith-Miles, Automatic age estimation based on facial aging patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007) 2234–2240.
[55] S. Goswami, S. Sarkar, M. Rustagi, Stylometric analysis of bloggers' age and gender, Proceedings of the Third International ICWSM Conference, 2009, pp. 214–217, (San Jose, CA) Association for the advancement of artificial intelligence. www.aaai.org.
[56] C. Peersman, W. Daelemans, L. Van Vaerenbergh, Predicting age and gender in online social networks, Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents — SMUC '11, ACM Press, New York, New York, USA, 2011, pp. 37–44.
[57] R.R. Prasath, Learning age and gender using co-occurrence of non-dictionary words from stylistic variations, RSCTC10 Proceedings of the 7th International Conference on Rough Sets and Current Trends in Computing, Springer-Verlag, Berlin, 2010, pp. 544–550, (Warsaw, Poland).
[58] R. Sarawgi, K. Gajulapalli, Y. Choi, Gender attribution: tracing stylometric evidence beyond topic and genre, Proceedings of the Fifteenth Conference on Computational Natural Language Learning, Association for computational linguistics, Stroudsburg, PA, 2011, pp. 78–86, (Portland, Oregon).
[59] R.R. Singh, D.S. Tomar, Approaches for user profile Investigation in Orkut Social Network, International Journal of Computer Science and Information Security 6 (2009) 259–268.
[60] T.S.H. Teo, V.K.G. Lim, Gender differences in internet usage and task preferences, Behaviour & Information Technology 19 (2000) 283–295.
[61] D.I. Holmes, The evolution of stylometry in humanities scholarship, Literary and Linguistic Computing 13 (3) (1998) 111–117.
[62] S. Abe, et al., Mining personal experiences and opinions from Web documents, Web Intelligence and Agent Systems: An International Journal 9 (2011) 109–121.
[63] H. Becker, M. Naaman, L. Gravano, Learning similarity metrics for event identification in social media, Conference on Web Search and Data Mining, ACM, New York, 2010, pp. 291–300, (New York).
[64] T. Fukuhara, H. Nakagawa, T. Nishida, Understanding sentiment of people from news articles: temporal sentiment analysis of social events, Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2007, (Boulder, Colorado) www.icwsm.org/2007.
[65] K. Inui, et al., Experience mining: building a large-scale database of personal experiences and opinions from web documents, International Conference on Web Intelligence and Intelligent Agent Technology, Ieee, Sydney, Austriala, 2008, pp. 314–321.
[66] L.-W. Ku, Y.-T. Liang, H.-H. Chen, Opinion extraction, summarization and tracking in news and blog corpora, Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006, pp. 100–107, (Menlo Park, CA) http://www.aaai.org/Library/Symposia/Spring/ss06-03.php.
[67] J. Landmann, C. Zuell, Identifying events using computer-assisted text analysis, Social Science Computer Review 26 (2007) 483–497.
[68] B.D. Longueville, R.S. Smith, G. Luraschi, "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, Proceedings of the 2009 International Workshop on Location Based Social Networks, ACM, New York, 2009, pp. 73–80, (Pittsburgh, PA).
[69] Q. Mei, C. Zhai, Discovering evolutionary theme patterns from text: an exploration of temporal text mining, International Conference on Knowledge Discovery and Data Mining, 2005, pp. 198–207, (Chicago, Ill).
[70] Q. Miao, Q. Li, R. Dai, AMAZING: a sentiment mining and retrieval system, Expert Systems with Applications 36 (2009) 7192–7198.
[71] H.-J. Min, J.C. Park, Identifying helpful reviews based on customer's mentions about experiences, Expert Systems with Applications 39 (2012) 11830–11838.
[72] K.C. Park, Y. Jeong, S.H. Myaeng, Detecting experiences from weblogs, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, Association for computational linguistics, Stroudsburg, PA, 2010, pp. 1464–1472, (Uppsala, Sweden).
[73] R. Saurí, J. Pustejovsky, Are you sure that this happened? Assessing the factuality degree of events in text, Computational Linguistics 38 (2) (2012) 261–299.
[74] B. Tsolmon, A.-R. Kwon, K.-S. Lee, Extracting social events based on timeline and sentiment analysis in Twitter corpus, in: G. Bouma (Ed.), Natural Language Processing and Information Systems, Springer, Berlin, 2012, pp. 265–270.
[75] T. Warren Liao, Clustering of time series data—a survey, Pattern Recognition 38 (2005) 1857–1874.
[76] B. Zhao, et al., Discovering collective viewpoints on micro-blogging events based on community and temporal aspects, Advanced Data Mining, Springer, Berlin/Heidelberg, 2011, pp. 270–284.
[77] Y. Zhao, Y. Zhao, K. Helsen, Consumer learning in a turbulent market environment: modeling consumer choice dynamics after a product-harm crisis, Journal of Marketing Research 48 (2) (2011) 255–267.
[78] D.R. Jones, C.a. Thompson, Identifying events using similarity and context, Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, Association for Computational Linguistics, Morristown, NJ, USA, 2003, pp. 135–141.
[79] B. Tsolmon, A.-R. Kwon, K.-S. Lee, Extracting social events based on timeline and sentiment analysis in twitter corpus, Natural Language Processing and Information Systems, Springer, Berlin/Heidelberg, 2012, pp. 265–270.
[80] E.V. Balaguer, P. Rosso, Detection of near-duplicate user generated contents: the SMS spam collection, Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents, ACM, New York, 2011, pp. 27–33, (Glasgow, UK).
[81] R. Chandy, H. Gu, Identifying spam in the iOS app store, Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality — WebQuality '12, ACM Press, New York, New York, USA, 2012, p. 56.
[82] T. Fawcett, F. Provost, Combining data mining and machine learning for effective user profiling, International Conference on Knowledge Discovery and Data Mining, 1996, pp. 8–13, (Portland, Oregon) AAAI. http://www.aaai.org/Papers/KDD/1996/KDD96-002.pdf.

[83] N. Hu, et al., Manipulation in digital word-of-mouth: a reality check for book reviews, Decision Support Systems 50 (2011) 627–635.
[84] E.-P. Lim, et al., Detecting product review spammers using rating behaviors, Proceedings of the 19th ACM International Conference on Information and Knowledge Management — CIKM '10, ACM Press, New York, New York, USA, 2010, p. 939.
[85] Y. Lu, et al., Exploiting social context for review quality prediction, Proceedings of the 19th International Conference on World Wide Web — WWW '10, ACM Press, New York, New York, USA, 2010, p. 691.
[86] A. Mukherjee, B. Liu, N. Glance, Spotting fake reviewer groups in consumer reviews, Proceedings of the 21st International Conference on World Wide Web, ACM, 2012, pp. 191–200.
[87] S. Xie, et al., Review spam detection via temporal pattern discovery, Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining — KDD '12, ACM Press, New York, New York, USA, 2012, p. 823.
[88] A.Z. Broder, Identifying and filtering near-duplicate documents, in: R. Giancarlo, D. Sankoff (Eds.), Combinatorial Pattern Matching, Springer-Verlag, Berlin/Heidelberg, 2000, pp. 1–10.
[89] D. Perito, et al., How unique and traceable are usernames? Privacy Enhancing Technologies, Springer-Verlag, Berlin Heidelberg, 2011, pp. 1–17.
[90] G. Adomavicius, A. Tuzhilin, User profiling in personalization applications through rule discovery and validation, International Conference on Knowledge Discovery and Data Mining, 1999, pp. 377–381.

**Fons Wijnhoven** has a master of science in Research Methodology from Radboud University Nijmegen and a PhD on Management Information Systems from Twente University Enschede. Fons is associate professor of knowledge management and information systems. His research is on information service business models, internet information quality and information credibility. He has a book on information services (Taylor and Francis 2012) and foundations of informing (Routledge 2009) and he has articles on information credibility and value in journals like Management Learning, Knowledge Management Research and Practice, and the Journal of the American Society on IST.

**Oscar Bloemen** has a master science degree in Physics and a master of science in Business Administration both from Twente University. Oscar is a consultant and developer of tools for sentiment mining and he owns a data science startup consultancy company. Oscar also has an article on physics in the journal of Fluid Dynamics.