"I GRADE WHAT I GET BUT WRITE WHAT I THINK." INCONSISTENCY ANALYSIS IN PATIENTS' REVIEWS

Complete Research

Geierhos, Michaela, University of Paderborn, Paderborn, Germany, geierhos@hni.upb.de
Bäumer, Frederik S., University of Paderborn, Paderborn, Germany, fbaeumer@hni.upb.de
Schulze, Sabine, University of Paderborn, Paderborn, Germany, sabine.schulze@hni.upb.de
Stuß, Valentina, University of Paderborn, Paderborn, Germany, valentina.stuss@hni.upb.de

Abstract

Received medical services are increasingly discussed and recommended on physician rating websites (PRWs). The reviews and ratings on these platforms are valuable sources of information for patient opinion mining. In this paper, we have tackled three issues that come along with inconsistency analysis on PRWs: (1) Natural language processing of user-generated reviews, (2) the disagreement in polarity of review text and its corresponding numerical ratings (individual inconsistency) and (3) the differences in patients' rating behavior for the same service category (e.g. 'treatment') expressed by varying grades on the entire data set (collective inconsistency). Thus, the basic idea is first to identify relevant opinion phrases that describe service categories and to determine their polarity. Subsequently, the particular phrase has to be assigned to its corresponding numerical rating category before checking the (dis-)agreement of polarity values. For this purpose, several local grammars for the pattern-based analysis as well as domain-specific dictionaries for the recognition of entities, aspects and polarity were applied on 593,633 physician reviews from both German PRWs jameda.de and docinsider.de. Furthermore, our research contributes to content quality improvement of PRWs because we provide a technique to detect inconsistent reviews that could be ignored for the computation of average ratings.

Keywords: Text-Rating-Inconsistency (TRI), Physician rating websites (PRW), Polarity inference.

1 Introduction

Due to the fact that Web 2.0 basically decreased the cost of communication, patients can compare medical services without effort, evaluate the performance of physicians and their staffs and share their personal experiences with the health care system (Cambria and Hussain, 2012). Hence, the number of physician-rating websites (PRWs)[1] is rising rapidly (Sabin, 2013; Terlutter et al., 2014) and patients are increasingly turning "to online physician ratings, just as they have sought ratings for other products and services" (Hanauer et al., 2014). These "online physician reviews are a massive and potentially rich source of information [for] capturing patient sentiment regarding healthcare" (Wallace et al., 2014). The comprehension of social information provided through the "Patient 2.0" is essential for the understanding of service quality and its improvement (Cambria and Hussain, 2012). The automatic analysis of this textual information, also known as sentiment analysis, has increasingly attracted attention in social media marketing and patient opinion mining (Cambria and Hussain, 2012). The advantage of opinion mining

[1] We identified the following current existing German PRWs in alphabetical order: arzt.weisse-liste.de, arzt-auskunft.de, aerzte-notdienst.de, bessereaerzte.de, docinsider.de, esando.de, imedo.de, jameda.de, medfuehrer.de, onmeda.de, sanego.de, topmedic.de, yourfirstmedicus.de.

on PRWs is that it avoids potential cognitive bias problems of 'traditional' elicitation techniques such as "social desirability" (Verhoef et al., 2014).

1.1 Problem Statement

However, the automatic analysis of patient reviews is a challenging task. For instance, there is the natural language processing (NLP) challenge of narrative comments additionally provided to numerical ratings on PRWs (Emmert et al., 2014). These textual evaluations give patients the opportunity to "elaborate upon their rating with additional comments and provide personal experiences and impressions that cannot be covered by the scaled rating system. [...] Additionally, the information provided can help physicians gain a better understanding of patient concerns" (Emmert et al., 2014). Anyway, it is not only (1) the natural language processing of the review texts itself that is addressed here, but also (2) the inconsistency between the polarity of narrative comments and the polarity of the corresponding numerical ratings (*individual inconsistency*) within a review. An example therefore is 'The treatment was very good!' in combination with a bad grade in the rating category 'treatment'. Besides, a further issue that has to faced here is (3) the individual rating behavior due to patients' expectations on a special service category (e.g. 'treatment') expressed through the different assigned grades for the same service quality (e.g. 'very good') on the entire data set (*collective inconsistency*). Possible explanations for such inconsistencies in review texts and ratings are that in "most of the practical cases [...] users do not carefully rate items, they either forget to rate the given items (i.e., missing value problem) or they make a mistake on the precise evaluation (i.e. noisy rating problem)" (Pham and Jung, 2013). Furthermore, "ratings are influenced by subjective factors (from user) and objective factors (from system) together. While user factors include various psychological effects, e.g., attitude, mentality, and satisfaction, the system factors are trust, interests, user interfaces and so on" (Pham and Jung, 2013).

1.2 Objectives

In this paper, we introduce a method to detect these inconsistencies for PRWs' and for users' benefit. When numerical ratings and review texts do not fit together, this "can be frustrating and [...] can reduce the value of the online review system." (Mudambi et al., 2014). Our findings help the PRWs to "automatically remove these comments and/or exclude them from the computation of average ratings. Alternatively, [they could] [...] alert users about the inconsistencies. Removing these [...] reviews also [...] reduce[s] the noise in the data set, yielding better performance in later analysis" (Fu et al., 2013). Thus, the basic idea is first to identify relevant opinion phrases that describe service categories and to determine their polarity. Subsequently, the particular phrase has to be assigned to its corresponding numerical rating category before checking the (dis-)agreement of polarity values. For this purpose, several local grammars for the syntactic analysis as well as domain-specific dictionaries for the recognition of entities, aspects and polarity were used. Therefore, our approach is special in that it enriches current research by evaluating the content of user-generated reviews on PRWs because other studies mainly focused on the numerical rating results so far (Emmert et al., 2014). The paper is structured as follows: The next chapter provides an overview on the related work. It is followed by the research design in Chapter 3 which introduces the data set (3.1), describes the data preprocessing (3.2) and presents the inconsistency analysis (3.3). In Chapter 4, we evaluate our approach and discuss our results before we conclude in Chapter 5.

2 Related Work

In this chapter, we present background information relevant to our work. We also survey related works and point out their relationship to our approach. First, we discuss approaches dealing with individual inconsistency in Section 2.1. Then we contrast with studies on collective inconsistency in Section 2.2.

2.1 Individual Inconsistency

"It is generally assumed that ratings are a numeric representation of text sentiments and their valences are consistent. This however may not always be true" (Hu et al., 2014). But if the numeric ratings differ from the review texts, this can be confusing. Islam (2014) therefor propose a "unified rating system" in order to resolve possible inconsistencies. For this purpose, they infer sentiment from texts to generate numeric ratings based on the polarity of the full-text review. The overall rating is then defined as the "average of the rating done by sentiment analysis and the star rating given by the users" (Islam, 2014). However, we do not distill numerical ratings from the texts in order to provide consistent reviews. We aim at detecting such inconsistencies by comparing the polarity of the text to the users' grades or stars.

"Ratings that do not match the actual text of the comments" (Fu et al., 2013) are, for example, discovered by WisCom. Although Fu et al. (2013) also analyze inconsistencies in reviews, they concentrate on mobile App marketplaces such as Google Play Store. However, these reviews are somehow different from physician reviews because they "are generally shorter in length since a large portion of them are submitted from mobile devices on which typing is not easy" (Fu et al., 2013). Unlike Fu et al. (2013) we do not apply a regression model either for the detection of inconsistencies between ratings and review texts. And yet others (e.g. Mudambi et al., 2014) try to detect misalignment between review texts and star ratings in order to understand why and where such inconsistencies are most likely to occur. In particular, they use a machine learning approach to determine if certain texts are specific for product reviews of n stars on amazon.com. Thus, the "classification algorithm can identify a 'typical' 5-star review, 3-star review, and so on" (Mudambi et al., 2014). Jang et al. (2014) also infer sentiment scores from texts in order to face the issue of "the inconsistency between textual evaluation (review content) and scoring evaluation (review rating)". However, they conduct their survey on hotel reviews. Since we focus on a different domain (German PRWs), we cannot apply machine learning techniques because only few physician reviews of the entire data set are affected by individual inconsistency (cf. Table 3).

Furthermore, Lak and Turetken (2014) investigate whether "sentiment analysis results can be used as an alternative to star ratings". Even though they also match the sentiments of reviews to the corresponding ratings as we do, there are some notable differences. First of all, they perform this comparison using cross tabulation and chi square analysis as well as a two-tailed bivariate correlation analysis. Secondly, they conduct part-of-speech tagging and then apply a sentiment analysis tool called "Lexicalytics" which returns "one single specific score in 3 decimal places between -1 to +1". However, we agree with Islam (2014) that part-of-speech tagging and parsing are not feasible for the analysis of review texts because of their informal writing style. Lak and Turetken (2014) study reviews of the same domain but not of the same language because RateMDS is an English- and not a German-speaking PRW. Moreover, they investigate just 50 comments on average for three general practitioners (from Toronto). In contrast to Lak and Turetken (2014), we analyze a more representative and larger data set of German physician reviews (cf. Section 3.1).

2.2 Collective Inconsistency

As there " exists no function for aggregating consistent individual sets of judgments over multiple interconnected propositions into consistent collective ones" (List, 2005), collective inconsistency appears in user-generated reviews. For instance, somebody's "view of a '3' may be considerable different from another's" (Mudambi et al., 2014). Furthermore, the rating might be inconsistent due to the fact that "mapping opinions to a single number is complex [...] for users" (Centeno et al., 2014). In particular, Maciejovsky and Budescu (2013) show that numerical and verbal recommendation formats (i.e. verbal reviews and numerical ratings) are "processed differently by consumers and can lead to different patterns of preferences. Thus, one can expect increased levels of preference inconsistency across the two modes" (Maciejovsky and Budescu, 2013). There have been a few studies that have attempted to find explanations for this "phenomenon" as evidenced by people who "post negative comments, even alongside a 5-star

product evaluation" (Mudambi et al., 2014). Possible reasons therefor are altruism, community-building motivations and reciprocity (Schau et al., 2009). An incentive for posting negative comments is that these are conceived as "more intelligent, competent and expert" (Amabile, 1983) than positive reviews.
Several studies investigated that someone else's reviews could influence one's rating behavior. Wu and Huberman (2008), for example, discovered that the "exposure of previous opinions to potential reviewers induces a trend following process which leads to the expression of increasingly extreme views". Talwar et al. (2007) confirmed the dependency of (numerical) ratings on previously read reviews. They figure out that "previous reports create an expectation of quality which affects the subjective perception of the user" (Talwar et al., 2007). According to McGlohon et al. (2010) "groups of users who amply discuss a certain feature are more likely to agree on a common rating for that feature." Even Gilbert and Karahalios (2010) observe that about 10 to 15 % "of [100,000 Amazon] reviews substantially resemble previous ones" because many reviewers just echo what previous users said.
Even before the rise of Web 2.0, "considerable heterogeneity in consumer interpretation and use of scales" (Mudambi et al., 2014) as well as the resulting response biases were widely discussed topics in opinion polling. There have been numerous studies on the design of surveys with respect to response scales (Weathers et al., 2005) and the biases or variety in response styles that arise due to the scale formats used (Greenleaf, 1992; Kieruj and Moors, 2010; Weijters et al., 2010). These studies discovered that different formats are perceived differently in terms of meaning and salience of response categories and thus have an influence on which response option on the scale is selected (Arce-Ferrer, 2006; Schaeffer and Presser, 2003). Furthermore, it was empirically proved that there is a "tendency [...] to choose the extreme endpoints of a rating scale" and "to make disproportionate use of the middle response category" (Kieruj and Moors, 2010). That is an observation we also made while analyzing collective inconsistency in online patients' reviews and therefore decided to group the extreme endpoints of the German grading system (1 & 2 and 5 & 6) in Section 3.3.2.
However, our goal is not to investigate the various reasons for such rating inconsistencies but to automatically detect if the same inconsistencies occurring in online product reviews or in traditional surveys can also be observed across all user-generated physician reviews.

3 Research Design

In this chapter, we describe the data collection and preprocessing as well as our approach towards inconsistency analysis. For this purpose, we created an extensive data set of physician reviews (cf. Section 3.1), developed domain-specific local grammars for information extraction and sentiment analysis (cf. Section 3.2) and describe how to uncover inconsistencies in physician reviews in Section 3.3.

3.1 Data Set

A data set is usually designed for a particular purpose. Because of no available ready-made corpora of physician reviews, we have to build our own specialized corpus. For corpus creation, we collected texts from two different German PRWs: jameda.de and docinsider.de. Both sites provide enough data for a balanced corpus because of their user popularity and traffic volume and their great amount of physician reviews. Furthermore, many medical subjects are covered (representativeness). By jameda.de's own account, it provides details on 250,000 physicians and five million reviews. More than 90 % of German physicians are registered on this PRW[2]. With more than 3.07 million users per month (2014), it is a well-known and widely accepted portal[3]. On docinsider.de we discovered 320,948 physicians, medical practices, clinics and other related institutions but the total number of reviews is unknown. In 2013 docinsider.de was the second-largest PRW in reach (behind jameda)[4].

[2] jameda GmbH, "FAQ", http://www.jameda.de/hilfe/?show=user, 11/19/14.
[3] Arbeitsgemeinschaft Online Forschung e.V., "internet facts 2014-08", http://www.agof.de/aktuelle-studie-internet/, 11/19/14.
[4] jameda GmbH, "GfK-Ranking: Reichweite von Arztbewertungsportalen", http://www.jameda.de/hilfe/?show=user, 11/19/14.

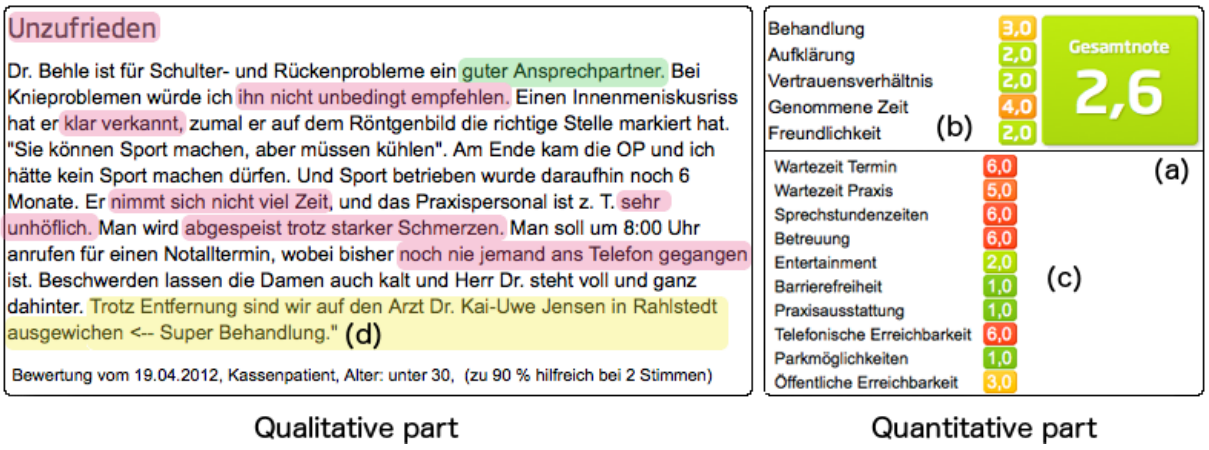

Figure 1. Sample review with labeled inconsistencies taken from jameda.de

By gathering data from jameda.de and docinsider.de between October 2013 and March 2014, we built a specialized corpus containing 593,633 individual physician reviews in total, where each review consists of a qualitative and a quantitative part (cf. Figure 1). While the textual information includes the title, review text and metadata (e.g. patient's personal data), there are also up to 16 numeric rating criteria (e.g. quality of treatment, equipment, organization). All reviews are written in German and were not edited, i.e. no spell checking or the like was applied. This data set covers the time period from January 2009 to December 2013, where the average length of a review text is 51 words and the longest one consists of 348 words. The descriptive metadata (e.g. age, type of health insurance) attached to each review provide classificatory information for a better understanding of the reviewer's background. In total, 61 % of all physician reviews contain details about the statutory health insurance (SHI) or the private medical insurance (PMI). Besides, 63 % provide information about the reviewer's age distributed in three main categories: 'younger than 30', 'between 30 and 50' and 'older than 50'. One criterion on which we will choose samples is the original data source. Because only five rating categories on jameda.de and six categories on docinsider.de are mandatory fields, the awarded number of rating criteria differs per review. Moreover, jameda.de uses a grading system (best: 1.0 – worst: 6.0) and docinsider.de applies star rating (best: 5 stars – worst: 1 star).

3.2 Data Preprocessing

Figure 2 illustrates the general workflow of our approach. First of all, information extraction is performed (Grishman, 1997). We therefore apply a method that automatically identifies and extracts relevant information from user-generated physician reviews and transforms it into a structured representation (i.e. predefined templates, Neumann, 2009). What should be recognized is defined by domain-specific rules (i.e. local grammars according to Nagel, 2010). So our information extraction system works with predefined patterns for entity recognition and aspect identification.

In our case, relevant expressions for the enrichment of the dictionaries and relevant phrases for the creation of local grammars are determined by frequency analysis on the physician reviews. We therefore generated n-grams (up to 5-grams) and build frequency lists. The most frequent n-grams of the length 5 are then the seed list for the construction of local grammars.

For instance, when the opinion phrase 'Dr. Foo is incompetent' occurs in the reviews, then a local grammar is able to detect the entity ('Dr. Foo'), the aspect ('incompetent') and substitute the adjective 'incompetent' by a variety of other negative adjectives such as 'unfriendly' or 'disinterested' belonging to the same category (here: 'treatment') of the given service category ontology (cf. Table 1). These words are taken from our lexical resources, especially the polarity dictionary SentiWS (Remus et al., 2010) and our own dictionaries (cf. Section 3.2.1) containing expressions that have already been identified in previous

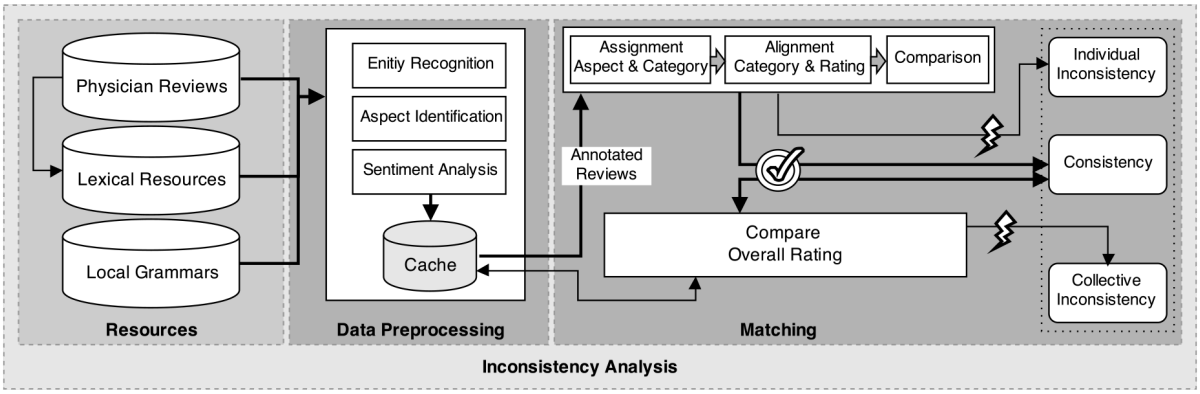

Figure 2. Inconsistency analysis workflow

preprocessing iterations by n-grams on our data set. In general, information extraction is an iterative process that is able to identify relevant opinion phrases in our data set by means of dictionary and local grammar usage (cf. Section 3.2.1 and 3.2.2).
After identifying the relevant opinion phrases (i.e. entities and aspects) from the data set, they are assigned to their corresponding textual polarity and a subcategory of our service category ontology. That is, for example, the positive opinion phrase 'The parking situation is great' is linked to the subcategory 'parking'. These subcategories are grouped into suitable main categories (e.g. 'tangibility' for the parking situation) in order to minimize the complexity of the alignment to the corresponding ratings in the course of the inconsistency analysis. Table 1 shows these subcategories for the service categories on jameda.de as well as some sample opinion phrases and their majority-aligned polarities (positive (P) or negative (N)) by the patients.
Finally, the polarity of each opinion phrase is compared to the numerical rating of the particular review in order to detect individual inconsistency. The consistent reviews with their respective ratings are then compared with each other in order to detect collective inconsistencies because even if the single reviews are (individual) consistent, the comparison of the reviews can reveal divergence in the rating behavior across users (i.e. collective inconsistent).

3.2.1 Lexical Resources

Lexical resources are crucial in most NLP tasks. Imagine, for instance, the following review text:

> *'The doctor is quite time-efficient. Without beating around the bush and with a mischievous smile, he told me that I have incurable leukemia. Then he said goodbye and went on to the next patient. All in all, a quite efficient patient handling!'*

Its polarity is unclear. Is this really a positive review? When placing an order on amazon.com for example, an 'efficient handling' is desirable. Also 'not beating around the bush' is preferable during a business meeting. But during consultation between physician and patient, less straightforwardness and time-efficiency are appreciated, especially for the receiver of bad news. Hence, almost nothing is entirely positive or negative (Olsher, 2012).

Since existing polarity lexical resources for German (e.g., SentiWS, Remus et al., 2010) do not reflect domain-specific lexical usage, we have to enrich these dictionaries and use context information for polarity classification. SentiWS "contains several positive and negative polarity bearing words weighted within the interval of [-1; 1] plus their part of speech tag, and if applicable, their inflections" (Remus et al., 2010).

category	subcategory	sample opinion phrase	polarity
assurance	child friendliness	*'Leseecke für Kinder'* ('reading corner for children')	P
	kindness	*'super netter Arzt'* ('extremely nice doctor')	P
	trust	*'absolut vertrauenswürdig'* ('absolutely trustworthy')	P
reliability	complementary medicine	*'keine alternativen Heilmethoden'* ('no alternative health practices')	N
	health education	*'fachlich gut beraten'* ('advised competently')	P
	treatment	*'kompetente Behandlung'* ('competent treatment')	P
responsiveness	consultation hours	*'super Öffnungszeiten'* ('super opening hours')	P
	public accessibility	*'fahren wenig Busse'* ('only reached by few buses')	N
	parking	*'keine Parkplätze'* ('no parking facilities')	N
	telephone accessibility	*'schlechte telefonische Erreichbarkeit'* ('poor accessibility by telephone')	N
	waiting time (appointment)	*'bekommt schnell einen Termin'* ('it is easy to get an appointment')	P
	waiting time (practice)	*'ewig lange Wartezeiten trotz Termin'* ('long waiting times despite appointment')	N
tangibility	accessibility	*'zuviele Stufen'* ('too many stairs')	N
	entertainment	*'kaum Zeitungen vorhanden'* ('almost no journals in the waiting room')	N
	practice equipment	*'tolle Praxis'* ('nice practice')	P
time	time taken	*'nimmt sich nur wenig Zeit'* ('only takes little time')	N

Table 1. Service category ontology

For our purpose, especially adjectives from SentiWS play a significant role. Moreover, we added the most frequently used adjectives in our data set to the corresponding lists for positive (P) and negative (N) uni- and bigrams. Furthermore, we use pattern dictionaries for the recognition of evaluative expressions (see bi- and trigrams in Table 2) and gazetteers work as specialized dictionaries (e.g. 48,046 doctor's names) to support initial tagging. On top of this, 1,377 domain-specific (medical) terms for diagnoses, syndromes and treatment were collected and organized in their own dictionary. This is necessary for the sentiment analysis in order to assign aspects to their corresponding entities (from medical terminology). Table 2 gives an overview on the created lexical resources which are grouped by patterns (n-grams) and polarity.

Neither these dictionaries nor their extensions can cover all variants of relevant opinion phrases. We therefore have to abstract from literal expressions and create semantically enriched syntactic patterns represented by local grammars. An example can be *'<A> Praxis'*, where <A> refers to an adjective in the dictionary such as *'unordentlich'* ('untidy') or other adjectives in the left context of *'Praxis'* ('practice').

n-gram	description	example	polarity	amount
n=1	single adjective	*'kompetent'* ('competent')	P	33,327
n=1	single adjective	*'hektisch'* ('hectic')	N	25,081
n=2/n=3	adjective phrase	*'sauber und modern'* ('clean and modern')	P	1,373
n=2/n=3	adjective phrase	*'sehr unordentlich'* ('very untidy')	N	297

Table 2. Polarity lexical resources for sentiment analysis

3.2.2 Local Grammars

Local grammars describe semantic-syntactic structures that cannot be formalized in electronic dictionaries. They are represented by directed acyclic graphs (cf. Figure 3) and implemented as finite state transducers (FST, Geierhos, 2010). These transducers produce output in terms of semantic annotations (i.e. labels) for recognized rating categories and evaluative expressions in the review texts. The grammar rules were instantiated with high-frequent n-grams and then generalized. For each of the five categories in Table 1 (assurance, reliability, responsiveness, tangibility and time), a local grammar was developed based on pattern dictionaries.

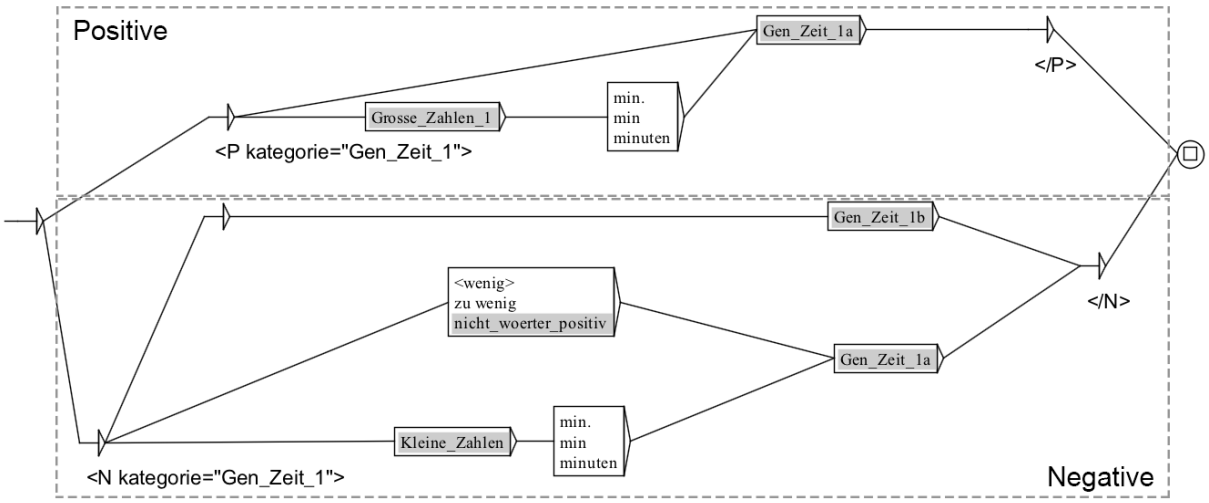

Figure 3. Finite state transducer for treatment time

Figure 3 shows the subgraph representing the FST for the recognition of treatment time ('Gen_Zeit_1') which belongs to the local grammar for 'time'. There are two main ways through this graph: While a short treatment time is annotated as negative (<N>) by the lower path, a long treatment time is labeled as positive (<P>) by the upper path of this local grammar. In particular, the detection of rather short treatment times is realized by the lower part of the FST by including characteristic words, e.g. *'wenig'* ('few'), and numbers expressing only few time. For this purpose, two subgraphs are imported which contain small numbers (e.g. "Kleine_Zahlen") and keywords which are often used to express a short period of time in this particular domain (e.g. "nicht_woerter_positiv"). Words like *'wenig'* ('few') which are surrounded by angle brackets are simultaneously looked-up in the corresponding list of inflected forms in the lexical resources, e.g. *'wenige'* ('few'), *'weniger'* ('fewer'), etc.

3.3 Inconsistency Analysis

We analyze two types of inconsistency in physician reviews. While the individual inconsistency (3.3.1) is restricted to a single review, the collective inconsistency (3.3.2) occurs across the whole data set due to patients' subjective rating behavior and thus requires a different methodological approach.

3.3.1 Individual Inconsistency

"Individual inconsistency is directly related to the assumption of random individual error" (Whitely, 1978). We therefore define individual inconsistencies in the course of this work as divergent polarities for each of the five categories (assurance, reliability, responsiveness, tangibility and time; cf. Table 1) in the qualitative and quantitative part of the same review. For this reason, we consider the disagreement of good ratings per category (grades 1 to 3 and 4 to 5 stars respectively) and negative patient's opinion in the review text as well as the mismatch of bad ratings (grades 4 to 6 and 1 to 3 stars respectively) and positive patient's statements as individual inconsistencies.

Figure 1 shows a sample review from the data set introduced in Section 3.1. It is divided in qualitative and a quantitative part. Furthermore, the overall score (a) is calculated as the average grade of the five mandatory rating categories (b). The voluntary ratings (c) are not considered for the arithmetic averaging. This example contains a lot of individual inconsistency occurrences. First, the review is entitled *'Unzufrieden'* ('dissatisfied') which indicates a bad rating. While the title expresses negative sentiment, the average grade of 2.6 on a scale that ranges from 1.0 (best) to 6.0 (worst) still implies satisfaction. Moreover, the first sentence of the review text refers to a positive past experience with this physician. It is therefore likely that the scores represent a mixture of several visits to the physician, which is a possible explanation for the following inconsistencies:

Opinion phrases such as *'ihn nicht unbedingt weiterempfehlen'* ('I cannot necessarily recommend him') and *'klar verkannt'* ('clearly misjudged') indicate the disturbance of trust and poor treatment. But the grade 2.0 for the trust in the physician-patient relationship expresses satisfaction and the treatment is graded quite positive (3.0), too. Referring to our definition of individual inconsistency, we observed divergent polarities. In the category 'reliability', there is a disagreement of the strongly negative statements *'ihn nicht unbedingt weiterempfehlen'* ('not necessarily recommend him') combined with the grade 2.0 for 'trust' and *'klar verkannt'* ('clearly misjudged') combined with the grade 3.0 for 'treatment' (cf. Table 1). Even more clearly is the contradiction between the given grade 2.0 for *'Freundlichkeit'* ('kindness') as subcategory of 'assurance' and the obviously negative remark *'sehr unhöflich'* ('very rude').

category	**subcategories**	**occurrence of individual inconsistency**
assurance	kindness, trust, ...	4.50 %
reliability	health education, ...	11.20 %
responsiveness	waiting time, ...	12.03 %
tangibility	entertainment, ...	8.41 %
time	time taken	2.95 %

Table 3. Empirical probability of individual inconsistencies occurring in the data set

In order to provide more detailed information about how often individual inconsistencies appear in the whole data set, we assigned all sentiment scores per review text to their corresponding rating categories (e.g. time taken, health education, kindness) in order to detect divergences. Then, we calculated the arithmetic mean for each category (assurance, reliability, responsiveness, tangibility and time) based on the identified individual inconsistencies per subcategory (cf. Table 1). Table 3 outlines their relative frequency per rating category in order to show which categories are more (individual) error-prone (i.e.

noisy). It is noticeable that the category 'responsiveness' is worst affected (12.03 %) while 'assurance' only contains 4.5 % inconsistent reviews.

3.3.2 Collective Inconsistency

The analysis of collective inconsistencies is based on the polarity of opinion phrases and their assigned grades across the whole data set. This kind of inconsistency occurs when an identical statement appears in various review texts and is not always aligned to the same grade (due to different reviewers). For this purpose, all local-grammar-annotated phrases were extracted from the data set and ranked in terms of frequency. Then the top-ten opinion phrases of each subcategory were compared to the corresponding grades. Table 4 shows a sample phrase of negative polarity and its corresponding ratings over all reviews for the subcategory 'time taken' that spread over four more or less bad grades (3 to 6).

category	**subcategory**	**opinion phrase**	**number of occurrences per grade**			
time	time taken	*'nimmt sich wenig Zeit'* ('takes little time')	*Grade 3*	*Grade 4*	*Grade 5*	*Grade 6*
			2	8	39	33

Table 4. Example pattern with grades and number of respective occurrences

Although the grades 1 and 2 do not exist in this example, we in general pair grades (1 and 2, 3 and 4, 5 and 6) per opinion phrase. While the majority of patients using this statement in their reviews are consistent when assigning 5 or 6 to the category 'time', there are ten who decided to give better grades (3 and 4). That way, inconsistencies can appear when people have different rating behaviors.
Collective inconsistencies can be observed for opinion phrases of positive or negative polarity, but the probability to occur differs among the various categories (cf. Table 5).

category	**occurrence of collective inconsistency**
assurance	22.90 %
reliability	26.55 %
responsiveness	33.87 %
tangibility	26.57 %
time	42.52 %

Table 5. Empirical probability of collective inconsistencies occurring in the data set

4 Evaluation

We evaluated the reliability of our approach and therefore present our evaluation methodology in the next section. In Section 4.2, we provide our evaluation results of the individual inconsistency analysis and the ones for the recognition of collective inconsistency which are discussed in Section 4.3.

4.1 Evaluation Methodology

The recognition rate of the consistencies highly depends on the local-grammar-annotated phrases. For evaluation purposes, the data set introduced in Section 3.1 was split into a training set (66.6 % of the total size) and a test set (33.4 %). We therefor use traditional evaluation measures, i.e. precision, recall and the balanced F-score. While true positives are hits (i.e. correctly classified inconsistencies), false positives count the number of erroneously assumed inconsistencies. If our inconsistency analysis failed to realize the occurrence of inconsistency, we call this a false negative.

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \tag{1}$$

$$\text{recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \tag{2}$$

$$\text{F}_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3}$$

4.2 Evaluation Results

Due to the different characteristics of each type of inconsistency, the evaluations were conducted separately.

4.2.1 Individual Inconsistency

We assumed that it is a hard challenge to consider all random errors causing individual inconsistency and therefore expected bad results for our first test run. Nevertheless, we could show that it possible to automatically uncover inconsistencies within a patient review although there is evidence for improvement. In Table 6, the average precision per category is shown. Here the precision indicates the recognition rate of the (correct) polarity disagreement of grades and opinion phrases within a single review. While the precision value for 'responsiveness' is 50 %, only 3 % of all assumed inconsistency occurrences in the category 'reliability' were correctly identified. Moreover, all existing occurrences of individual inconsistency in the test set were found over all categories. Thus, the recall is 100 % per each category.

category	**precision**	**recall**	**F_1-score**
assurance	8 %	100 %	15 %
reliability	3 %	100 %	6 %
responsiveness	50 %	100 %	67 %
tangibility	28 %	100 %	44 %
time	14 %	100 %	25 %

Table 6. Evaluation results of individual inconsistency analysis

4.2.2 Collective Inconsistency

For evaluation purposes, only local-grammar-annotated opinion phrases were considered. Since inflection in language produces a great variety of forms, e.g. *'schlechter'* ('worse'), *'schlechtester'* ('worst'), for the same words, e.g. *'schlecht'* ('bad'), we decided to group similar opinion phrases into pattern represented by regular expressions (e.g. *'schlecht*'*). To analyze the quality of our approach for the detection of collective inconsistencies, we randomly selected a pattern for each category (e.g. *'schlecht* arzt'* ('bad / worse / worst doctor') for the category 'assurance') and compared the different grades therefor given. Table 7 shows our promising results.

4.3 Discussion

The strengths and weaknesses of the proposed approach are analyzed as follows. On the one hand, it was shown that it is possible to find inconsistencies by means of a local grammar-based analysis of opinion phrases. Local grammars perfectly fit our needs because of their domain-specific character and problem-modularity. The evaluation results show that our preprocessing step in Section 3.2 has the same shortcomings as other pattern-based approaches: The reliability of our inconsistency analysis depends on the predefined patterns for opinion phrases to be recognized. Once a pattern becomes too general within

category	pattern	precision	recall	F_1-score
assurance	*'schlecht* arzt'* ('bad / worse / worst doctor')	89 %	76 %	82 %
reliability	*'(\w+)+ schlecht behandelt'* ('badly treated')	77 %	63 %	69 %
responsiveness	*'mit etwas wartezeit'* ('with a bit waiting time')	100 %	93 %	96 %
tangibility	*'komisch* praxis'* ('strange practice')	100 %	86 %	92 %
time	*'wenig zeit genommen'* ('took only little time')	100 %	93 %	96 %

Table 7. Evaluation results of collective inconsistency analysis

a local grammar, too many phrases will match this pattern (i.e. overgeneration) – even some refer to past visits or other physicians named within the same review (see (d) in Figure 1) – and will be assigned to a false category. However, too specific patterns reduce the recall because we cannot conduct any inconsistency analysis on opinion phrases we did not identify during the preprocessing (i.e. overfitting). In user-generated reviews, we have to deal with lots of misspellings. Then, no approximate matching is possible with local grammars and therefore opinion phrases with spelling errors are ignored. Another issue is the assignment of wrong polarities and categories. For example, *'Ich musste bisher nie <N kategorie="WZP">lange auf einen Termin warten</N>'* ('So far I never had to wait that long for an appointment') is a positive statement for waiting time but was annotated with <N> (negative) during the preprocessing because the word *'nie'* ('never') in the left context shifting the polarity of *'lange'* ('long') had not been recognized. Although our approach achieves quite good results (cf. Table 7), it still needs improvement in terms of more precise patterns and classification accuracy. Furthermore, some opinion phrase patterns can belong to more than one category of the five defined as assurance, reliability, responsiveness, tangibility and time. This means that, for example, *'komisch* Praxis'* ('strange practice') can be assigned to 'tangibility' because it is related to the facilities of a practice or to 'assurance' when it describes the patient's overall feeling about the practice and the treatment.

5 Conclusion and Future Work

The increasing amount of patients that share their personal experiences with the health care system on PRWs generates a large amount of valuable information that is also error-prone.

5.1 Lessons Learned

In this paper, we have tackled three issues that come along with inconsistency analysis on PRWs: (1) Natural language processing of user-generated reviews, (2) the disagreement in polarity of review text and its corresponding numerical ratings (*individual inconsistency*) and (3) the differences in patients' rating behavior for the same service category (e.g. 'treatment') expressed by varying grades on the entire data set (*collective inconsistency*).

Since such contradictions are annoying, frustrating and confusing for patients seeking for information on PRWs, we developed an approach to detect these inconsistencies. The basic idea was first to identify relevant opinion phrases and to determine their polarity. Subsequently, the particular phrase was assigned to its corresponding category (assurance, reliability, responsiveness, tangibility and time) and then aligned to its grade before checking the (dis-)agreement of both (textual and numerical) polarity values. For this purpose, several local grammars for the pattern-based analysis as well as domain-specific dictionaries

for the recognition of entities, aspects and polarity were applied on 593,633 physician reviews from both German PRWs jameda.de and docinsider.de. Thus, our approach is special in that it enriches current research by evaluating the content of user-generated reviews on PRWs because it does not only focus on numerical ratings.

Our results show that collective as well as individual inconsistencies exist not only in product but also in physician reviews. We determined the empirical probability of individual and the collective inconsistencies to occur in different rating categories. While individual inconsistency dominates the category 'responsiveness' (12.03 %), the collective inconsistency mostly affects the category 'time' (42.52 %). To sum up, the collective inconsistency occurs more frequently (30.48 % on average) than the individual inconsistency (7.81 %) in our data set.

These results are useful to solve the rating inference problem. That is, when we know, for example, that for the majority of patients a waiting time of thirty minutes corresponds to the grade 3, then this is a valuable information for the interpretation of user-generated reviews or rating predictions respectively. Furthermore, our research contributes to content quality improvement of PRWs because we provide a technique to detect inconsistent reviews that could be ignored for the computation of average ratings.

5.2 Perspectives

However, our approach has to be further improved. First of all, our local grammars have to be extended to cover a greater variety of opinion phrases in future. Another challenge is the pattern enhancement – while manual refinement is promising but time-consuming, we have to experiment with several machine learning approaches for domain-specific pattern acquisition to avoid overfitting or overgeneration. Moreover, our approach is still monolingual. We intend to adapt it to further languages and will therefor create additional domain-specific dictionaries. A further fascinating task for future research is to resolve wrong assignments of patterns to specific categories. For instance, how to decide that the expression 'strange practice' belongs to the behavior of the staff and not to the practice facilities. Furthermore, we have to develop a strategy how to deal with off-topic opinion phrases referring to other physicians than the one that is on-topic in the review.

Acknowledgments

This work was supported in part by a research grant from the University of Paderborn and by a grant from the Ministry of Innovation, Higher Education and Research of North Rhine-Westphalia, Germany.

References

Amabile, Teresa M (1983). "Brilliant but cruel: Perceptions of negative evaluators." *Journal of Experimental Social Psychology* 19 (2), 146–156. ISSN: 0022-1031. DOI: 10.1016/0022-1031(83)90034-3.

Arce-Ferrer, Alvaro J (2006). "An Investigation Into the Factors Influencing Extreme-Response Style Improving Meaning of Translated and Culturally Adapted Rating Scales." *Educational and Psychological Measurement* 66 (3), 374–392. ISSN: 0013-1644. DOI: 10.1177/0013164405278575.

Cambria, Erik and Amir Hussain (2012). *Sentic Computing: Techniques, Tools, and Applications*. Vol. 2. SpringerBriefs in Cognitive Computation. Springer. ISBN: 9789400750708.

Centeno, Roberto et al. (2014). "On the inaccuracy of numerical ratings: dealing with biased opinions in social networks." *Information Systems Frontiers*, 1–17. ISSN: 1387-3326. DOI: 10.1007/s10796-014-9526-1.

Emmert, Martin et al. (2014). "What do patients say about their physicians? An analysis of 3000 narrative comments posted on a German physician rating website." *Health Policy* 118 (1), 66–73. DOI: 10.1016/j.healthpol.2014.04.015.

Fu, Bin et al. (2013). "Why people hate your app: Making sense of user feedback in a mobile app store." In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. KDD '13. ACM. Chicago, Illinois, USA, pp. 1276–1284. ISBN: 9781450321747. DOI: 10.1145/2487575.2488202.

Geierhos, Michaela (2010). *BiographIE – Klassifikation und Extraktion karrierespezifischer Informationen*. Vol. 5. Linguistic Resources for Natural Language Processing. Munich, Germany: Lincom. ISBN: 9783862880133.

Gilbert, Eric and Karrie Karahalios (2010). "Understanding Deja Reviewers." In: *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*. CSCW '10. ACM. Savannah, Georgia, USA, pp. 225–228. ISBN: 9781605587950. DOI: 10.1145/1718918.1718961.

Greenleaf, Eric A (1992). "Improving Rating Scale Measures by Detecting and Correcting Bias Components in Some Response Styles." *Journal of Marketing Research* 29 (2), 176–188. DOI: 10.2307/3172568.

Grishman, Ralph (1997). "Information extraction: Techniques and challenges." In: *Information Extraction A Multidisciplinary Approach to an Emerging Information Technology*. Springer, pp. 10–27.

Hanauer, David A et al. (2014). "Public Awareness, Perception, and Use of Online Physician Rating Sites." *JAMA* 311 (7), 734–735. DOI: 10.1001/jama.2013.283194.

Hu, Nan et al. (2014). "Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales." *Decision Support Systems* 57, 42–53. ISSN: 0167-9236. DOI: 10.1016/j.dss.2013.07.009.

Islam, Mir R (2014). "Numeric rating of Apps on Google Play Store by sentiment analysis on user reviews." In: *Proceedings of the International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*. IEEE. Dhaka, Bangladesh, pp. 1–4. ISBN: 9781479948208. DOI: 10.1109/ICEEICT.2014.6919058.

Jang, Wooseok et al. (2014). "Why the Online Customer Reviews Are Inconsistent? Textual Review vs. Scoring Review." In: *Digital Enterprise Design & Management*. Springer International Publishing, p. 151. ISBN: 9783319043128. DOI: 10.1007/978-3-319-04313-5_20.

Kieruj, Natalia D and Guy Moors (2010). "Variations in response style behavior by response scale format in attitude research." *International journal of public opinion research* 22 (3), 320–342. DOI: 10.1093/ijpor/edq001.

Lak, Parisa and Ozgur Turetken (2014). "Star Ratings versus Sentiment Analysis–A Comparison of Explicit and Implicit Measures of Opinions." In: *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS)*. IEEE. Waikoloa, Hawaii, USA, pp. 796–805. DOI: 10.1109/HICSS.2014.106.

List, Christian (2005). "The probability of inconsistencies in complex collective decisions." *Social Choice and Welfare* 24 (1), 3–32. ISSN: 0176-1714. DOI: 10.1007/s00355-003-0253-7.

Maciejovsky, Boris and David V Budescu (2013). "Verbal and numerical consumer recommendations: Switching between recommendation formats leads to preference inconsistencies." *Journal of Experimental Psychology: Applied* 19 (2), 143.

McGlohon, Mary et al. (2010). "Star Quality: Aggregating Reviews to Rank Products and Merchants." In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. Washington, D.C., USA: Association for the Advancement of Artificial Intelligence, pp. 114–121.

Mudambi, Susan M et al. (2014). "Why Aren't the Stars Aligned? An Analysis of Online Review Content and Star Ratings." In: *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS)*. IEEE. Waikoloa, Hawaii, USA, pp. 3139–3147. DOI: 10.1109/HICSS.2014.389.

Nagel, S. (2010). *Lokale Grammatiken zur Beschreibung von lokativen Sätzen und ihre Anwendung im Information Retrieval*. Vol. 12. Studien zur Informations- und Sprachverarbeitung. Munich, Germany: Centrum für Informations- und Sprachverarbeitung. ISBN: 9783930859283.

Neumann, Günter (2009). "Informationsextraktion." In: *Computerlinguistik und Sprachtechnologie. Eine Einführung*. Ed. by Kai-Uwe Carstensen et al. Spektrum Akademischer Verlag, pp. 594–605.

Olsher, Daniel J (2012). "Full spectrum opinion mining: Integrating domain, syntactic and lexical knowledge." In: *Proceedings of the 12th International Conference on Data Mining Workshops (ICDMW)*. IEEE. Brussels, Belgium, pp. 693–700. ISBN: 9781467351645. DOI: 10.1109/ICDMW.2012.166.

Pham, Hau X and Jason J Jung (2013). "Preference-based user rating correction process for interactive recommendation systems." *Multimedia tools and applications* 65 (1), 119–132. DOI: 10.1007/s11042-012-1119-8.

Remus, Uwe et al. (2010). "SentiWS – a Publicly Available German-language Resource for Sentiment Analysis." In: *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA), pp. 1168–1171. ISBN: 2951740867.

Sabin, James E (2013). "Physician-rating websites." *Virtual Mentor* 15 (11), 932–936.

Schaeffer, Nora C and Stanley Presser (2003). "The science of asking questions." *Annual Review of Sociology* 29, 65–88. DOI: 10.1146/annurev.soc.29.110702.110112.

Schau, Hope J et al. (2009). "How brand community practices create value." *Journal of Marketing* 73 (5), 30–51. DOI: 10.1509/jmkg.73.5.30.

Talwar, Arjun et al. (2007). "Understanding user behavior in online feedback reporting." In: *Proceedings of the 8th ACM conference on Electronic commerce*. EC '07. ACM. San Diego, California, USA, pp. 134–142. ISBN: 9781595936530. DOI: 10.1145/1250910.1250931.

Terlutter, Ralf et al. (2014). "Who Uses Physician-Rating Websites? Differences in Sociodemographic Variables, Psychographic Variables, and Health Status of Users and Nonusers of Physician-Rating Websites." *Journal of medical Internet research* 16 (3). DOI: 10.2196/jmir.3145.

Verhoef, Lise M et al. (2014). "Social Media and Rating Sites as Tools to Understanding Quality of Care: A Scoping Review." *Journal of medical Internet research* 16 (2). DOI: 10.2196/jmir.3024.

Wallace, Byron C et al. (2014). "A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews." *Journal of the American Medical Informatics Association* 21 (6), 1098–1103. DOI: 10.1136/amiajnl-2014-002711.

Weathers, Danny et al. (2005). "The impact of the number of scale points, dispositional factors and the status quo decision heuristic on scale reliability and response accuracy." *Journal of Business Research* 58 (11), 1516–1524. DOI: 10.1016/j.jbusres.2004.08.002.

Weijters, Bert et al. (2010). "The effect of rating scale format on response styles: The number of response categories and response category labels." *International Journal of Research in Marketing* 27 (3), 236–247. ISSN: 0167-8116. DOI: 10.1016/j.ijresmar.2010.02.004.

Whitely, Susan E (1978). "Individual inconsistency: Implications for test reliability and behavioral predictability." *Applied Psychological Measurement* 2 (4), 571–579. DOI: 10.1177/014662167800200412.

Wu, Fang and Bernardo A Huberman (2008). "How Public Opinion Forms." In: *Internet and Network Economics*. Vol. 5385. Springer, pp. 334–341. ISBN: 9783540921851. DOI: 10.1007/978-3-540-92185-1_39.