User Experience Mining auf Basis von Online-Produktbewertungen

David Lechler, Manuel Burghardt

Lehrstuhl für Medieninformatik, Universität Regensburg

Zusammenfassung

Die vorliegende Arbeit untersucht die Grenzen und Möglichkeiten einer automatischen Erfassung der *User Experience* (UX) von Produkten durch die Analyse von Online-Reviews. Hierzu wurde ein Tool entwickelt, das online verfügbare Produktbewertungen sammelt und mittels Methoden des *Natural Language Processing* sowie der Sentiment-Analyse aufbereitet, um die Ergebnisse auf die Werteskalen des *User Experience Questionnaire* (UEQ) abzubilden. Weiterhin präsentieren wir eine Evaluation des so erstellten Tools, indem wir die automatisch generierten Ergebnisse mit den Ergebnissen einer klassischen Benutzerstudie vergleichen. Es zeigt sich, dass automatisch erstellte UX-Analysen aus Online-Reviews zumindest eine grobe Annäherung zur UX eines Produkts zulassen – wenn auch nicht so differenziert wie ein von echten Anwendern vollständig ausgefüllter UEQ.

1 Einleitung und Fragestellung

Das *Social Web* bietet Nutzern die Möglichkeit ihre Meinung zu äußern und sich miteinander zu vernetzen. Dabei steht der Austausch von Meinungen und Informationen durch Kommentare, geteilte Inhalte und Bewertungen im Vordergrund (Chi, 2008, S. 88 f.). Im Zuge dessen bietet das *Social Web* in Form von Produkt-Reviews auch Informationen zum Nutzererlebnis, d. h. es finden sich detaillierte Bewertungen zu unterschiedlichen Dienstleistungen und Produkten, welche somit eine wertvolle Quelle zur Analyse der *User Experience* (UX) darstellen. Meist ist die Zahl der vorhandenen Reviews jedoch zu groß, um sie in angemessener Zeit ohne technische Hilfsmittel zu erfassen. Wir präsentieren deshalb einen Ansatz, der versucht mithilfe von Sentiment-Analyse-Techniken (Liu, 2012) online verfügbare Produkt-Reviews automatisch anhand der darin beschriebenen UX-Informationen auszuwerten. Die Sentiment-Analyse (SA), auch Opinion Mining genannt, analysiert Meinungen, Empfindungen, Bewertungen, Standpunkte und Emotionen bezogen auf Produkte, Dienstleistungen, Organisationen, Individuen sowie Ereignisse und deren Attribute. Insbesondere das Teilgebiet des S*entiment Mining* (Liu, 2012, S. 17), also die computergestützte Erfassung, Aufbereitung und Auswertung von nutzergenerierten Inhalten, scheint vielversprechend für die automatische Analyse

Veröffentlicht durch die Gesellschaft für Informatik e. V. 2017 in
M. Burghardt, R. Wimmer, C. Wolff, C. Womser-Hacker (Hrsg.):
Mensch und Computer 2017 – Tagungsband, 10.–13. September 2017, Regensburg.
 https://doi.org/10.18420/muc2017-mci-0228

von Produktreviews. In Analogie zum bestehenden *Sentiment Mining* und zum allgemeiner ausgerichteten *Text Mining* bezeichnen wir diesen Vorgang als *UX Mining*.

Ein kurzer Überblick über verwandte Studien zum Thema *UX Mining* findet sich in Kap. 2. Im weiteren Verlauf der Arbeit werden die methodische Vorgehensweise sowie die daraus resultierenden Erkenntnisse anhand bestehender Forschungsergebnisse evaluiert. In Kap. 3 präsentieren wir den *UxMiner*, ein Tool, das Produkt-Reviews sammelt und mithilfe von *Natural Language Processing* (NLP) und *Sentiment Analysis* UX-relevante Informationen extrahiert. Die Qualität solch automatisch gewonnener UX-Informationen wird in einer Evaluationsstudie in Kap. 4 systematisch überprüft und mit Ergebnissen einer parallelen UX-Nutzerstudie verglichen. Der Beitrag schließt mit einer Diskussion der Grenzen und Möglichkeiten des *UxMiners* (Kap. 5) und gibt einen Ausblick zu künftigen Optimierungsmöglichkeiten (Kap. 6).

2 Related work

Jijkoun et al. (2010) präsentieren einen Leitfaden zur Identifikation von Nutzererlebnissen in Online-Foren. Hedegaard & Simonsen (2013) untersuchen die Verteilung von Informationen zu Usability- und UX-Skalen in Online-Reviews und extrahieren daraus ein skalen-spezifisches Vokabular. Dabei finden sie heraus, dass 13 % – 49 % aller untersuchten Sätze der Reviews Informationen bezüglich Usability oder UX enthalten. Weiterhin stellen sie fest, dass die unterschiedlichen UX-Skalen je nach Produktkategorie unterschiedlich stark ausgeprägt sind und das Vokabular mitunter stark variieren kann bzw. häufig nur einige wenige Wörter umfasst (Hedegaard & Simonsen, 2013, S. 2093 ff.). Mendes et al. (2015) analysieren qualitativ Kommentare des sozialen Netzwerks *Twitter* und eines Universitäts-Netzwerkes und kommen zu dem Ergebnis, dass es möglich ist, bestimmte Äußerungen den Skalen der Usability oder UX zuzuordnen, bedeutende Nutzer-Probleme anhand dieser Äußerungen zu bestimmen und den zugehörigen Nutzungskontext zu ermitteln. Zu ähnlichen Resultaten kommen auch Oh & Lee (2015) durch eine manuelle Analyse von Online-Reviews hinsichtlich UX-Problemen bei der Anwendung von *quantified self trackern*. Ein domänen-spezifisches Lexikon wurde von Zhu & Fang (2014), basierend auf der Annahme, dass Reviews das Nutzererlebnis in natürlicher Sprache wiedergeben, erstellt. Durch die semi-automatische Faktorenanalyse hunderttausender Reviews zu Videospielen wurden sechs Faktoren zur Beschreibung von Videospielen ermittelt.

Die den genannten Studien zugrundeliegenden Einschränkungen hinsichtlich der untersuchten Domäne und des Ausdrucksvermögens der ermittelten UX legen nahe, dass weitere Forschungsarbeiten notwendig sind, um das Potenzial von Reviews und den darin enthaltenen UX-Informationen auszuschöpfen. In Anbetracht der Beobachtungen in diesen verwandten Studien muss davon ausgegangen werden, dass Informationen in Reviews hinsichtlich der UX mit einem Produkt je nach Domäne unterschiedlich geäußert werden. Inwiefern eine Zuordnung solch variierender Äußerungen zu übergreifenden Skalen der UX umgesetzt werden kann, soll im Rahmen dieser Studie analysiert werden. Ebenso soll untersucht werden, wie

sich die Ergebnisse einer automatischen Analyse von Produkt-Reviews zu einer klassischen UX-Nutzerstudie mit demselben Produkt als Untersuchungsgegenstand verhalten.

3 Design eines *UX-Mining*-Tools

In diesem Kapitel beschreiben wir das Design und die Umsetzung des *UxMiners*, einem Tool, das in der Lage ist, Produkt-Reviews der englischsprachigen Online-Plattform *Amazon.com* automatisch anhand der darin beschriebenen UX-Informationen zu analysieren. Die so ermittelten Ergebnisse werden dann automatisch auf eine bestehende UX-Metrik, den *User Experience Questionnaire* (UEQ) übertragen. Der UEQ ist aufgrund der empirischen Auswahl seiner semantischen Differenziale und der inhärenten Reliabilität und Konstruktvalidität (Laugwitz et al., 2008, S. 73) ein Standard-Werkzeug der UX-Evaluation.

3.1 Systemarchitektur des *UxMiners*

Der *UxMiner* wurde mithilfe von Python und unter Rückgriff auf bestehende Bibliotheken entwickelt. Insbesondere das für die Verarbeitung natürlicher Sprache verfügbare Python *Natural Language Toolkit* (NLTK[1]) wurde für die sprachliche Analyse der Reviews genutzt. Die weiteren verwendeten Bibliotheken erleichtern zum einen das Extrahieren von Informationen aus HTML-Code (*Beautiful Soup 4*[2]) und zum anderen das Erstellen des abschließenden UX-Reports in Form einer Excel-Datei (*openpyxl*[3], *jdcal*[4] und *xmlfile*[5]). Der *UxMiner* gliedert sich in drei Module: den *UxCrawler*, den *UxProcessor* und den *UxMapper*.

UxCrawler – Die Reviews von *Amazon.com* werden zunächst vom *UxCrawler* per *http-requests* über die *Representational State Transfer*-Schnittstelle (REST) angefordert. Auf die Verwendung der bestehenden *Amazon API* wurde verzichtet, da hier Reviews nicht direkt abgerufen werden können[6]. Der *UxCrawler* liefert die folgenden Informationen zurück und speichert sie in einer CSV-Datei: Titel, Text, „Hilfreich"-Bewertung des Reviews, „Sterne"-Bewertung des Review-Verfassers, Datum, Nutzername, Review-URL, Review-ID, Angabe zum verifizierten Kauf, die insgesamt erhaltenen „Hilfreich"-Bewertungen zu allen Reviews des Verfassers und der von *amazon.com* ermittelte Reviewer-Rang unter allen Review-Schreibern.

UxProcessor – Der *UxProcessor* bereitet Titel und Texte der übergebenen Reviews für deren abschließende Analyse mithilfe von Python NLTK auf. Zu Beginn werden aus dem Titel und dem Text eines jeden Reviews Stoppwörter entfernt. NLTK bietet hierzu eine Liste mit englischen Stoppwörtern, die von uns noch weiter angepasst wurde. Insbesondere wurden negie-

[1] http://www.nltk.org/; Anmerkung: Alle URLs in diesem Artikel wurden zuletzt am 24.3.2017 aufgerufen.
[2] https://www.crummy.com/software/BeautifulSoup/.
[3] https://openpyxl.readthedocs.io/.
[4] https://pypi.python.org/pypi/jdcal.
[5] https://pypi.python.org/pypi/et_xmlfile.
[6] https://docs.aws.amazon.com/AWSECommerceService/latest/DG/EX_RetrievingCustomerReviews.html.

rende Wörter und einige bedeutungsverstärkende Wörter entfernt. Weiterhin werden die Reviews in einzelne Tokens segmentiert und automatisch anhand der jeweils identifizierten Wortarten annotiert. Abschließend werden alle Wörter lemmatisiert, d. h. auf ihre Grundform zurückgeführt. Die Ergebnisse des *UxProcessors* werden ebenfalls in einer CSV-Datei abgespeichert, welche dann vom *UxMapper* weiterverarbeitet werden kann.

UxMapper – Aufgabe des *UxMapper*-Moduls ist der Abgleich (*mapping*) zwischen den Wörtern der Review-Texte und den semantischen Differenzialen des UEQ. Beim Abgleich werden einige sprachliche Herausforderungen berücksichtigt, bspw. Synonym-Erkennung über *WordNet* (Boyd-Graber et al., 2006) sowie auch die Berücksichtigung von Negation, Bedeutungsverstärkung und Komparativ oder Superlativ. Beim Mapping wurden zudem Synonyme für alle UEQ-Differenziale in den Reviews mithilfe des NLTK-Moduls *WordNet* identifiziert (Agirre & Edmonds, 2006).

3.2 Berechnung der UEQ-Werte

UEQ-Differenziale besitzen numerische Werte entsprechend einer Likert-Skala aus dem Wertebereich W = [1; 7]. Die Wörter der UEQ-Differenziale nehmen dabei die Extrempositionen 1 und 7 ein. Ausgehend von der neutralen Position 4 verändert sich bei entsprechender Nennung eines Wortes der Wert eines Differenzials in die entsprechende Richtung. Mehrfachnennungen eines Wortes verändern dessen UEQ-Wertigkeit jeweils um den Wert 1. Die Steigerungsformen Komparativ und Superlativ wurden ebenfalls bei der Berechnung der Wertigkeiten der Differenziale mit einbezogen. Je nach vorliegender Ausprägung addieren oder subtrahieren Komparative den Wert 2 und Superlative den Wert 3. Die Extremwerte der UEQ-Skala, also 1 und 7, können dabei jedoch nie über- bzw. unterschritten werden.

Tritt die Verneinung eines Worts aus den UEQ-Differenzialen auf, wird im Sinne der Bedeutungsumkehrung zum jeweiligen Gegensatzpaar im Differenzial gewechselt. Ein Lexikon mit gewichteten Bedeutungsverstärkern wurde von Brooke (2009) übernommen. Der prozentuale Wert eines Bedeutungsverstärkers wird dabei mit der semantischen Orientierung (SO) eines Worts multipliziert und zum ursprünglichen Wortwert hinzuaddiert. Negation kehrt das Vorzeichen des Bedeutungsverstärkers entsprechend um[7].

3.3 Ergebnisse des *UxMappers*

Der *UxMapper* erstellt eine CSV-Datei, die neben den aufbereiteten Daten auch die berechneten UEQ-Werte der Wortpaare enthält. Diese CSV-Datei kann problemlos gefiltert werden und ermöglicht dem Benutzer individuelle Anpassungen. So können beispielsweise sehr kurze oder ältere Reviews von der Analyse ausgenommen werden. Da trotz Filtermöglichkeit sehr umfangreiche Datenkollektionen zu erwarten sind, wurde die standardmäßige UEQ-Obergrenze von 1.000 auf maximal 10.000 Datensätze erhöht. Entsprechend wurden auch die be-

[7] Berechnungsbeispiel: *somewhat* (negative Bedeutungsverstärkung: -30%); *sleazy* (SO: -3). Somit ergibt sich für den Ausdruck “somewhat sleazy” folgender SO-Wert: -3 + (-3 * -30%) = -2 (vgl. Brooke, 2009, S. 31).

rücksichtigten Bereiche der UEQ-Formeln zur Berechnung des Mittelwerts, der Standardabweichung etc. angepasst, um die darauf aufbauende Berechnung der Konfidenzintervalle nicht zu verfälschen.

4 Evaluationsstudie

Dieses Kapitel beschreibt eine Evaluationsstudie zum Vergleich der automatisch vom *UxMiner* generierten Ergebnisse mit den Ergebnissen einer klassischen Nutzerstudie. Die Nutzerstudie wurde dabei als Online-Umfrage zum Produktbereich „Smartphone“ durchgeführt. Die Umfrage wurde von 238 Teilnehmern im Zeitraum einer Woche (21. - 28. Januar 2016) bearbeitet. Die Teilnehmer wurden einerseits nach ihrem aktuell verwendeten Smartphone befragt und sollten andererseits ihr bisheriges Nutzererlebnis mit dem Gerät in Form eines UEQ-Fragebogens dokumentieren. Die UEQ-Angaben zu den beiden in dieser Umfrage am häufigsten genannten Smartphones, dem *iPhone 5s* (17 Nennungen) und dem *Samsung Galaxy S4 Mini* (zwölf Nennungen), wurden als Vergleichsbasis für die automatischen Ergebnisse des *UxMiners* ausgewählt.

4.1 *UxMiner*-Analyseergebnisse für das *iPhone 5s*

Insgesamt wurden 2.330 verfügbare Reviews (60.955 Wörter) analysiert. Dabei wurden 1.286 Übereinstimmungen mit UEQ-Begriffen identifiziert. Dies entspricht durchschnittlich 0,55 Treffern pro Review. Die Anzahl dieser Treffer ist aufgrund von Mehrfachnennungen von UEQ-Differenzialen innerhalb eines Reviews größer als die insgesamt erfassten Bewertungen der UEQ-Wortpaare (951). Hieraus ergibt sich, dass im Durchschnitt lediglich 0,4 unterschiedliche UEQ-Items pro Review bewertet werden konnten. Dabei verdeutlichen diese Zahlen einen hohen Anteil fehlender Werte je Review.

Für sieben der 26 Differenziale konnten gar keine Übereinstimmungen in den Reviews gefunden werden. Die Differenziale „*easy to learn – difficult to learn“*, „*not interesting – interesting“* und „*unpredictable – predictable“* erzielten jeweils nur eine Übereinstimmung, weshalb die Berechnung der Varianz und der Standardabweichung hier fehlschlug. Nur den Items der UEQ-Skala *Attractiveness* konnten je mindestens drei Treffer zugeordnet werden. Alle verbleibenden Skalen besitzen mindestens ein nicht bewertetes Differenzial. Einige der Wortpaare haben besonders hohe Trefferzahlen: „*good – bad“* (*Attractiveness*) mit 595 Treffern, „*fast – slow“* (*Efficiency*) mit 129 Treffern, „*meets expectations – does not meet expectations“* (*Dependability*) mit 91 Treffern, “*complicated – easy”* (*Perspicuity*) mit 53 Treffern und “*unlikable – pleasing”* (*Attractiveness*) mit 29 Treffern. Die restlichen Items erzielten maximal zehn Übereinstimmungen. Die Mittelwerte der fünf genannten Wortpaare mit hohen Trefferzahlen liegen zwischen 0,8 und 1,5. Die Mittelwerte der verbleibenden Differenziale liegen zwischen -1,0 und 1,0. Die Standardabweichungen aller Items (0,0 bis 1,2) weisen keine auffälligen Ausreißer auf.

Die Mittelwerte der Skalen *Attractiveness* (M = 1,09), *Perspicuity* (M = 1,18), *Efficiency* (M = 0,98), *Dependability* (M = 0,82) und *Stimulation* (M = 1,0) liegen nah beieinander im posi-

tiven Bereich. Lediglich die Skala *Novelty* weicht hiervon mit einem Mittelwert von 0,50 deutlich ab. Das größte Konfidenzintervall liegt bei der Skala *Novelty* vor (-0,48 bis 1,48), wohingegen die Konfidenzintervalle der Skalen *Attractiveness* (1,0 bis 1,17), *Perspicuity* (0,91 bis 1,44), *Efficiency* (0,79 bis 1,16) und *Dependability* (0,66 bis 0,98) enger ausfallen.

4.2 *UxMiner*-Analyseergebnisse für das *S4 Mini*

Insgesamt wurden 2.338 verfügbare Reviews (57.924 Wörter) für das *Samsung Galaxy S4 Mini* analysiert. Dabei wurden 1.321 Übereinstimmungen mit UEQ-Differenzialen erkannt. Durchschnittlich konnten 0,56 Übereinstimmungen pro Review erfasst werden. Aufgrund der Mehrfachnennung einzelner Differenziale innerhalb einzelner Reviews fällt auch bei der Analyse des *S4 Mini* die Anzahl der insgesamt abgegebenen Bewertungen zu UEQ-Wortpaaren geringer aus (1.021). Dies entspricht einem Durchschnitt von 0,43 bewerteten UEQ-Differenzialen pro Review.

Für sechs der UEQ-Differenziale fanden sich gar keine Übereinstimmungen in den Reviews. Die drei meist genannten Wortpaare sind „*complicated – easy"* (82 Treffer), „*fast – slow"* (146 Treffer) und „*good – bad"* (621 Treffer). Mit 24 Treffern hebt sich zusätzlich das Wortpaar „*unlikable – pleasing"* von den übrigen Items durch die Trefferanzahl ab. Diese Differenziale sowie die meisten anderen Differenziale wurden im Mittel positiv bewertet (0,8 bis 1,5). Auffällig sind je zwei Items mit neutralen beziehungsweise negativen Mittelwerten. Diese sind "*easy to learn – difficult to learn"* (M = 0,3), "*usual – leading edge"* (M = -0,3), "*creative – dull"* (M = -1,0) und "*annoying – enjoyable"* (M = -1,3).

Die positivsten Mittelwerte weisen die Skalen *Attractiveness* (M = 1,11), *Perspicuity* (M = 1,06) und *Stimulation* (M = 1,000) auf. Ebenfalls positive Tendenzen weisen die Skalen *Efficiency* (M = 0,83) und *Dependability* (M = 0,80) vor. Mit einem Mittelwert von -0,38 zeigt die Skala *Novelty* als einzige eine neutrale Bewertung. Aufgrund der geringen Trefferanzahl der Skala *Stimulation* und der Skala *Novelty* müssen deren Resultate kritisch betrachtet werden.

4.3 Diskussion der *UxMiner*-Analysen

Der *UxMiner* konnte für das *iPhone 5s* 19 von 26 Wortpaaren mindestens einmal anhand der Reviews bewerten. Beim *S4 Mini* konnten für sechs Items gar keine Übereinstimmungen gefunden werden. Besonders markant ist, dass dieselben fünf Differenziale bei beiden Evaluationsgegenständen am häufigsten bewertet wurden. Das Item „*good – bad"* ist dabei mit 595 Treffern beim *iPhone 5s* und 621 Übereinstimmungen beim *S4 Mini* besonders auffällig. Entsprechend den ähnlichen Trefferzahlen der Items beider Analysen fallen auch die Trefferzahlen pro Skala relativ ähnlich aus. *Attractiveness* erzielt in beiden Analysen vor *Efficiency* die meisten Übereinstimmungen, was der Tatsache geschuldet ist, dass das Wortpaar „*good – bad"* aus dieser Skala stammt. In der Mitte der Trefferanzahlen liegen die Skalen *Perspicuity* und *Dependability*, wohingegen die Skalen *Stimulation* und *Novelty* in beiden Analysen mit Abstand am wenigsten Treffer generieren. Die Anzahl der Treffer pro Wortpaar und pro Skala korrelieren möglicherweise mit der gewählten Produktkategorie, jedoch kann man annehmen,

dass einige Items generell öfter von Review-Verfassern verwendet werden als andere. Insbesondere beim Item „*good – bad*“ erscheint dieser Gedanke nachvollziehbar, handelt es sich hierbei doch um ein grundlegendes, vielfältig einsetzbares Gegensatzpaar.

Bei den zwei durchgeführten *UxMiner*-Analysen konnten pro Review im Durchschnitt weniger als 0,5 UEQ-Differenziale bewertet werden. Lässt man allerdings die Anzahl der Reviews ohne Übereinstimmungen mit den UEQ-Differenzialen (~66 %, jeweils beim *iPhone 5s* und beim *S4 Mini*) außer Acht, so ergibt sich ein anderes Bild: 785 Reviews führten beim *iPhone 5s* zu 951 Bewertungen von UEQ-Items (1,21 pro Review). Beim *S4 Mini* erzielten 792 Reviews insgesamt 1021 Wertungen von UEQ-Differenzialen (1,28 pro Review). Die Reviews ohne Treffer des *iPhone 5s* weisen eine durchschnittliche Textlänge von 14,41 Wörtern und 3,12 Titelwörtern vor. Die Review-Texte mit UEQ-Übereinstimmungen des *iPhone 5s* sind durchschnittlich 39,13 Wörter lang, die Review-Titel bestehen im Schnitt aus 3,70 Wörtern. Auch beim *S4 Mini* zeigen sich für Reviews ohne Treffer (12,53 Textwörter; 2,83 Titelwörter) und für Reviews mit Treffern (39,0 Textwörter; 3,86 Titelwörter) ähnliche Durchschnittswerte.

4.4 Vergleich der UEQ-Daten der Online-Umfrage mit den automatischen Ergebnissen des *UxMiners*

Wie in Abbildung 1 zu sehen ist, zeigt sich beim *iPhone 5s* auf den ersten Blick ein deutlicher Unterschied in der Höhe der Werte, vergleicht man die Skalenmittelwerte der Online-Umfrage (*OU Mean*) mit denen der *UxMiner*-Analyse (*UxMiner Mean*). Auch die Häufigkeit der Bewertungen innerhalb der UEQ-Skalen liegt erwartungsgemäß deutlich auseinander (*OU N* und *UxMiner N*). Die Skalenmittelwerte der *UxMiner*-Analyse liegen deutlich unter denen der Online-Umfrage, weisen jedoch eine höhere Standardabweichung (*SD*) auf. Bei genauerer Betrachtung zeigt sich aber eine ähnliche Rangordnung, wenn man die Skalenmittelwerte in absteigender Reihenfolge ordnet. Festzuhalten ist, dass die Skalenmittelwerte der *UxMiner*-Analyse deutlich weniger stark ausgeprägt sind als die der Online-Umfrage.

Scales	OU Mean	OU N	OU SD	UxMiner Mean	UxMiner N	UxMiner SD
Attractiveness	1,971	17	0,654	1,090	644	1,083
Perspicuity	2,353	17	0,538	1,175	64	1,086
Efficiency	1,721	17	0,661	0,978	135	1,065
Dependability	1,868	17	0,553	0,819	95	0,816
Stimulation	1,206	17	0,746	1,000	9	0,000
Novelty	0,794	17	1,196	0,500	4	1,000

Abbildung 1: UEQ-Skalenmittelwerte und Standardabweichung der Online-Umfrage (OU) zum iPhone 5s und der UxMiner-Analyse.

Ein Vergleich der Skalenmittelwerte (*Mean*), die bewerteten Items der Skalen (*N*) und die Standardabweichung (*SD*) des *S4 Mini* aus den Ergebnissen der Online-Umfrage (OU) und der *UxMiner*-Analyse kann Abbildung 2 entnommen werden. Zwar besitzen alle Skalen in beiden Fällen dieselben Vorzeichen, bei genauerer Betrachtung jedoch werden einige Besonderheiten deutlich. Zum einen liegen die Skalenmittelwerte der Online-Umfrage zum *S4 Mini* näher an denen der *UxMiner*-Analyse. Zum anderen zeigen zwei Skalen (*Efficiency* und *Stimulation*)

der *UxMiner*-Analyse einen positiver ausgeprägten Mittelwert als ihre Gegenstücke aus der Online-Umfrage.

Scales	OU Mean	OU N	OU SD	UxMiner Mean	UxMiner N	UxMiner SD
Attractiveness	1,375	12	0,384	1,113	665	1,036
Perspicuity	1,500	12	0,853	1,060	94	0,811
Efficiency	0,542	12	0,789	0,829	158	1,004
Dependability	1,250	12	0,612	0,800	85	0,842
Stimulation	0,458	12	0,562	1,000	10	0,000
Novelty	-0,417	12	1,046	-0,375	9	0,916

Abbildung 2: UEQ-Skalenmittelwerte und Standardabweichung der Online-Umfrage (OU) zum S4 Mini und der UxMiner-Analyse.

Zur Beurteilung der Güte der Werte der *UxMiner*-Analysen werden stets die UEQ-Werte der Online-Umfrage als Goldstandard herangezogen. Auffällig ist, dass die Skalenmittelwerte der *UxMiner*-Analyse mit zwei Ausnahmen geringer ausfallen. Diese Tatsache lässt sich durch das ausgewogenere Bewertungsspektrum eines Reviews in Form von berücksichtigten Mehrfachnennungen einzelner UEQ-Differentiale (insbesondere gegensätzlicher Bewertungen) begründen. Hieraus ergibt sich ein enger gefasstes Bewertungsverhalten, weshalb seltener die Extrempositionen eines UEQ-Wortpaares eingenommen werden, was zu höheren Skalenmittelwerten führen würde.

Die Skalen *Stimulation* und *Novelty* zeigten in beiden Analysen die geringste Anzahl an Übereinstimmungen, weshalb deren Werte kritisch zu betrachten sind. In Anbetracht der vorliegenden Verteilung der zugehörigen Item-Bewertungen muss von einer unzureichenden Ausgewogenheit der Wortpaare dieser zwei Skalen für die Produktkategorie „Smartphones“ ausgegangen werden. Die Skalenmittelwerte der restlichen Skalen weisen zwar eine mit der Online-Umfrage übereinstimmende Tendenz auf, müssen jedoch als unzuverlässig betrachtet werden, da alle Skalen von wenigen Wortpaaren geprägt werden. Somit können die Skalenmittelwerte keinen ausgewogenen Gesamteindruck widerspiegeln.

5 Allgemeine Diskussion

Die vorliegende Studie macht deutlich, dass automatisch erstellte UX-Analysen aus Online-Reviews ein großes Potenzial haben. Gleichzeitig bestehen eine Reihe bekannter NLP- und Sentiment-Analyse-Probleme, die es künftig zu lösen gilt. Insbesondere der fehlerhafte Umgang mit Negation in Reviews kann enormen Einfluss auf den resultierenden UEQ nehmen. Weiterhin muss an dieser Stelle nochmals betont werden, dass der *UxMiner* ungeachtet eventuell vorliegender, abweichender Bezüge, *alle* Informationen aus den Reviews als produktbezogen interpretiert und entsprechend in die Bewertung des UEQs einfließen lässt.

Die Ergebnisse des Tools müssen auch deshalb vorsichtig gedeutet werden, da ihnen keine vollständigen UEQ-Datensätze zugrunde liegen. Dies bedeutet zum einen, dass statistische

Kenngrößen wie Konfidenz, Skalenkonsistenz und Item-Korrelationen teilweise nicht berechnet werden können. Zum anderen führen die fehlenden UEQ-Differenziale zu unzuverlässigen Skalen-Ergebnissen. Skalenmittelwerte werden unter Umständen von einzelnen Items mit sehr großen Trefferzahlen dominiert, wodurch die UEQ-Skalen ihre eigentliche Aussagekraft verlieren. Diese Beobachtung deckt sich mit den Ergebnissen von Hedegaard & Simonsen (2013) bezüglich der Verwendung bestimmter Wörter in Abhängigkeit der Produktkategorie. Die starre Zuordnung einzelner semantischer Differenziale, wie etwa „*good – bad*“ (*Attractiveness*), scheint bei genauerer Betrachtung fehleranfällig zu sein, können doch mit dem beispielhaft genannten Differenzial auch Attribute bewertet werden, die eher einer anderen Skala zuzuordnen wären. Dieser Gedanke wirft gleichzeitig die Frage auf, wie eine alternative Zuordnung der Differenziale erfolgen und ob diese in jedem Fall eindeutig vollzogen werden kann. Solche Überlegungen überstrapazieren jedoch den UEQ nicht nur hinsichtlich der zugrundeliegenden Konstruktvalidität, sondern auch bezüglich der Grundidee des UEQs, ein vorgefertigtes Konstrukt aus Wortpaaren auf verschiedene Produkte anwenden zu können.

Die bis hierhin aufgezeigten Grenzen des *UxMiners* offenbaren zugleich dessen Potenzial: Häufig genannte UEQ-Differenziale ermöglichen es, unter Berücksichtigung ihrer verallgemeinerten Aussagekraft, Rückschlüsse aus der Menge an Reviews zu ziehen. So ist es denkbar, sollten die Bewertungen eines oft auftretenden Wortpaares zu einer eindeutigen Tendenz führen, dass Aussagen über die Richtung dieses Wortpaares getroffen werden können. Ein Beispiel hierfür wäre, ob ein Produkt eher den Erwartungen der Käufer entspricht oder nicht.

6 Fazit und Ausblick

In der Gesamtschau zeigt sich, dass das *UxMiner*-Tool aufgrund der nicht zufriedenstellenden Reliabilität der UEQ-Ergebnisse eine klassische UX-Studie nicht ersetzen kann. Die Beantwortung der Frage, ob sich mit automatisierten Methoden des *Sentiment Mining* Aussagen über die UX eines Produktes treffen lassen, muss jedoch differenzierter erfolgen: Einerseits können anhand häufig auftretender UEQ-Wortpaare nachvollziehbare Aussagen über das Produkt getroffen werden. Andererseits muss kritisch hinterfragt werden, inwieweit eine solche Schlussfolgerung gesichert ist. Alles in allem sind jedoch auch solche Aussagen aufgrund der angeführten Unsicherheiten nicht äquivalent zu den möglichen Schlussfolgerungen, die durch eine klassische UX-Studie in der Regel ermöglicht werden. Sie bieten lediglich Einblicke in die Produkt-UX, ohne auf die Ursachen hierfür eingehen zu können. Somit ist der *UxMiner* etwa für Vorab-Einschätzungen dienlich, um in weiteren UX-Studien bestimmte Schwerpunkte zu setzten oder einzelne Aspekte gezielt zu hinterfragen.

Abschließend muss auf den Einsatz vorhandener Verfahren zum Umgang mit den zahlreich auftretenden fehlenden Werten verwiesen werden, die potenziell zu einer weiteren Verbesserung der Ergebnisse führen können. In der Statistik finden sich hier zahlreiche Methoden, um fehlende Werte unter bestimmten Umständen anhand vorliegender Daten zu ergänzen, etwa die *Multiple Imputation*-Methode (Cheema, 2014) und der *Expectation-Maximization*-Algorithmus (Lüdtke & Robitzsch, 2010). Allerdings ist zu vermuten, dass keine zufällige Verteilung der fehlenden UEQ-Werte vorliegt (Lüdtke et al., 2007, S. 105). Diese Vermutung stützt sich auf die Annahme, dass verschiedene Produktkategorien aufgrund der Beschaffenheit der

zugehörigen Produkte unterschiedliche fehlende Werte in den UEQ-Skalen zur Folge haben. Inwiefern dieser Ansatz anwendbar ist und welche Güte die resultierenden Daten vorweisen, müssen weitere Studien zeigen.

Literaturverzeichnis

Agirre, E. & Edmonds, P. (2006). *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht: Springer.

Boyd-Graber, J., Fellbaum, C., Osherson, D. & Schapire, R. (2006). Adding dense, weighted connections to WordNet. In *Proceedings of the third international WordNet conference*, S. 29-36.

Brooke, J. (2009). *A Semantic Approach to Automated Text Sentiment Analysis*. Simon Fraser University. Online: https://www.sfu.ca/~mtaboada/docs/Julian_Brooke_MA_Sentiment_Analysis.pdf [03.03.2017].

Cheema, J. (2014). A Review of Missing Data Handling Methods in Educational Research. *Review of Educational Research, 84*(4), 487-508.

Chi, E. (2008). The Social Web: Research and Opportunities. *IEEE Computer 41*(9), 88-91. Retrieved from: http://www-users.cs.umn.edu/~echi/papers/2008-IEEE-Computer/socialweb-research.pdf [03.03.2017].

Hedegaard, S. & Simonsen, J. G. (2013). Extracting usability and user experience information from online user reviews. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, S. 2089-2098.

Jijkoun, V., de Rijke, M., Weerkamp, W., Ackermans, P. & Geleijnse, G. (2010). Mining user experiences from online forums: an exploration. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media, (June)*, S. 17-18.

Laugwitz, B., Held, T. & Schrepp, M. (2006). User Experience Questionnaire (English Version) [Data_Analysis_Excel_UEQ_English]. Retrieved from http://www.ueq-online.org/?slide=ueq-download.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies* (Vol. 5). San Rafael: Morgan & Claypool.

Lüdtke, O. & Robitzsch, A. (2010). Umgang mit fehlenden Daten in der empirischen Bildungsforschung. In S. Maschke & L. Stecher (Hrsg.), *Enzyklopädie Erziehungswissenschaft Online. Fachgebiet Methoden der empirischen erziehungswissenschaftlichen Forschung, Quantitative Forschungsmethoden*. Weinheim: Juventa.

Lüdtke, O., Robitzsch, A., Trautwein, U. & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. *Psychologische Rundschau, 58*(2), 103-117.

Mendes, M. S., Furtado, E., Furtado, V. & Castro, M. F. de (2015). Investigating Usability and User Experience from the user postings in Social Systems. In G. Meiselwitz (Hrsg.), *Social Computing and Social Media*. Los Angeles: Springer International Publishing, S. 216-228.

Oh, J. & Lee, U. (2015). Exploring User Experience Issues in Quantified Self Technologies. In *Proceedings of the Eighth International Conference on Mobile Computing and Ubiquitous Networking (ICMU)*, S. 53-59.

Zhu, M. & Fang, X. (2014). Introducing a Revised Lexical Approach to Study User Experience in Game Play by Analyzing Online Reviews. In *Proceedings of the 2014 Conference on Interactive Entertainment*, S. 1-8.