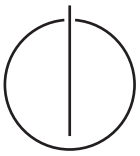
FAKULTÄT FÜR INFORMATIK

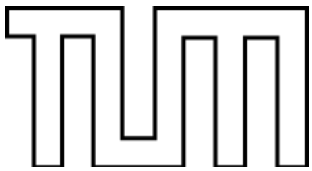
TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelorarbeit

Die Analyse der Stimmungen und Benutzerpersönlichkeiten in App Stores

Margarete Barth





FAKULTÄT FÜR INFORMATIK

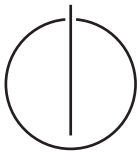
TECHNISCHE UNIVERSITÄT MÜNCHEN

Bachelorarbeit

Towards sentiment analysis and user personality studies in the App Store

Die Analyse der Stimmungen und Benutzerpersönlichkeiten in App Stores

Author:	Margarete Barth
Supervisor:	Prof. Bernd Brügge, Ph. D.
Advisor:	Emitzá Guzman
Submission Date:	15.08.2014



Ich versichere, dass ich diese Bachelor's Thesis selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

München, 15.08.2014 Margarete Barth

Acknowledgments

Abstract

Der Apple App Store bietet den Usern von Applikationen die Möglichkeit, diese mit Hilfe einer Sterne-Bewertung und eines schriftlichen Kommentars zu bewerten. Diese Studie untersucht, ob und wie der App Store als Kanal dienen kann, um Informationen über die User zu gewinnen. Durch die Analyse des User Feedbacks wird überprüft, ob es möglich ist, Erkenntnisse über die Persönlichkeitseigenschaften und Stimmungen der User zu generieren. Diese Informationen über die User der eigenen App oder ähnlicher Applikationen könnten dazu verwendet werden, die Anforderungserhebung oder das Beta-Testing bei der Wartung oder Neuentwicklung zu unterstützen. Es wird eine Vorgehensweise vorgestellt, mit Hilfe derer fünf exemplarische Forschungsfragen und deren zu Grunde liegenden Hypothesen beantwortet werden. Für die Erfassung der Persönlichkeiten und Stimmungen der User wird das text mining tool LIWC (Linguistic Inquiry and Word Count) verwendet, das alle Wörter des Feedbacks auf Zugehörigkeit zu verschiedenen sprachlichen und psychologischen Dimensionen untersucht. Durch die Verwendung statistischer Methoden konnten einige Erkenntnisse über die User acht verschiedener Applikationen aus fünf verschiedenen Ländern gewonnen werden. Es wurde festgestellt, dass sich die Eigenschaften der User abhängig von der verwendeten Applikation und der Zufriedenheit mit der Applikation stark von einander unterscheiden. Die Herkunft der User hatte nur bei manchen Applikationen einen Einfluss auf die Persönlichkeitseigenschaften der User. Außerdem wurde ermittelt, dass die Stimmungen, die in den Reviews ausgedrückt werden, mit den Sterne-Bewertungen positiv korrelieren und daher in weiteren Analysen als ein feineres Maß für die Zufriedenheit der User dienen können. Diese Studie kann als Road Map verwendet werden, um User Informationen aus dem User Feedback des Apple App Stores zu generieren.

Contents

Acknowledgments **v**

Abstract **ix**

1. Einführung **1**
1.1. App Stores als Kanal für die Charakterisierung der User 1
1.2. Motivation 3

2. Related Work **7**

3. Hypothesen **9**

4. Forschungsansatz **13**
4.1. App Store Daten 13
4.1.1. Generierung 13
4.1.2. Überblick über den Datensatz 15
4.2. Textdaten 18
4.2.1. Laden der Textdaten 18
4.2.2. Vorbereitung der Daten auf die Analyse 18
4.3. Metadaten 19
4.4. LIWC – Analyse der Textdaten 19
4.4.1. Linguistic Inquiry and Word Count Tool (LIWC) 19
4.4.2. Anwendung 21
4.5. Bestimmung der Persönlichkeitseigenschaften 21
4.5.1. Big Five Personality Traits Ansatz 21
4.5.2. Berechnung der Big Five 22
4.6. Berechnung der Sentiments 24
4.7. Statistische Analysen 24
4.7.1. Vortests 24
4.7.2. Kruskal-Wallis-Test 28
4.7.3. Wilcoxon-Rangsummen-Test 29
4.7.4. Spearmans Rangkorrelationskoeffizient 30

5. Ergebnisse **31**
5.1. Einfluss der Applikation auf die Eigenschaftswerte 32
5.2. Abhängigkeit der Big Five von dem Herkunftsland der Reviewer 33

5.3. Zusammenhang der Big Five Werte mit der Sterne Bewertung der User . 35
5.4. Korrelation der Sentiments mit den vergebenen Sternen 36
5.5. Veränderungen der User-Sentiments bei verschiedenen Versionen eine App 38

6. Interpretation und Future Work **41**
6.1. Applikationsabhängige Unterschiede 41
6.2. Abhängigkeit der Big 5 von der Herkunft der Reviewer 43
6.3. Einfluss der Sterne-Bewertungen auf Persönlichkeitseigenschaften 44
6.4. Positive Korrelation zwischen Sterne-Bewertungen und Sentiments 45
6.5. Abhängigkeit der Sentiments von der Versionsnummer 46

7. Conclusion **47**

A. Appendix **49**

Glossary **69**

Acronyms **71**

List of Figures **73**

List of Tables **75**

Bibliography **77**

1. Einführung

In diesem Kapitel werden zunächst Hintergrundinformationen gegeben, die letztendlich zu der Haupthypothese dieser Bachelorarbeit führen. Daraufhin werden die Ziele und Motivation zur Durchführung dieser Studie erklärt.

1.1. App Stores als Kanal für die Charakterisierung der User

Da der Apple App Store in dieser Studie eine wichtige Rolle spielt, werden über ihn zunächst ein paar Informationen gegeben. Der Apple App Store[1] wurde 2008 von Apple eröffnet und stellte damals ca. 500 Apps zur Verfügung (2013 waren es schon ca. 900 000 Stück). Mit Hilfe des App Stores sollte der Erwerb und die Verwaltung von Applikationen für das Apple Betriebssystem iOS erleichtert werden. Dieses erfolgreiche Konzept wurde in den darauffolgenden Jahren von anderen Anbietern übernommen (wie zum Beispiel Google, Windows usw.). Der Begriff „App Store“ wird mittlerweile allgemein für alle Plattformen mobiler Applikationen verstanden und nicht mehr ausschließlich für den Apple App Store verwendet. In den letzten Jahren stieg die Bedeutung von mobilen Applikationen und deren Vertriebsplattformen stark an, was in Tabelle 1.1 dargestellt wird. Die folgende Studie wird sich auf den Apple App Store konzentrieren, der neben dem Google Play Store[2] die meisten Applikationen anbietet und daher eine große Bedeutung hat. Außerdem stellt der App Store ausreichend Reviews für die Analysen zur Verfügung [BK11] [Sta].

Der Apple App Store bietet verschiedenste Applikationen für iPhones und iPads zum Download an. Einige davon kostenlos, andere kostenpflichtig. Nach dem Download einer App gibt es außerdem die Funktion eine App zu bewerten (Download ist jedoch Voraussetzung). Hierfür muss mindestens eine Sterne-Bewertung von einem bis fünf Sterne abgegeben werden (1 Stern = „Gefällt mir gar nicht.“, 5 Sterne = „Ist toll.“). Zusätzlich kann auch eine Rezension verfasst werden, in der man beschreibt, was man für Erfahrungen mit der App gemacht hat, was einem gut oder schlecht gefallen hat und auch eventuelle Anregungen an die Entwickler, wo Bugs oder Probleme aufgetreten sind. Dieses Feedback dient einerseits anderen potenziellen Kunden sich über die Vor- und Nachteile einer App zu informieren und andererseits auch den Entwicklern Verbesserungspotentiale der App zu erkennen. Das Schreiben von Rezensionen ist schon seit Eröffnung des App Stores im Juli 2008 möglich. Zwei Applikationen, die in dieser Studie analysiert werden, nämlich Evernote und Tripadvisor, gehören zu den

[1] https://itunes.apple.com/us/genre/ios/id36
[2] https://play.google.com/store?hl=de

	2009		2010		2011	
	Apps	Entwickler	Apps	Entwickler	Apps	Entwickler
Apple App Store	126.206	28.152	306.815	65.919	342.141	75.850
BlackBerry App World	4.412	1.110	17.923	2.660	27.029	3.927
Google Android Market	17.966	5.177	149.214	27.811	217.155	41.000
Nokia Ovi Store	6.556	638	25.150	3.396	31.023	4.642
Palm App Catalog	1.899	612	5.191	1.124	6.398	1.229
Windows Marketplace	873	357	6.779	3.068	13.522	5.170

Table 1.1.: Überblick über den Wachstum des App-Angebots und der aktiven Entwickler in den Jahren 2009-2011 [BK11]

ursprünglichen Applikationen des App Stores. Deren Rezensionen reichen bis zum Juli 2008 zurück [BI].

In dieser Studie wird überprüft, ob der App Store als Kanal dienen kann, um Informationen über die User verschiedener Applikationen zu offenbaren. Konkreter wird analysiert, ob es möglich ist an Hand des User Feedbacks, das die Kunden zu den Applikationen verfassen, Erkenntnisse über die Persönlichkeitseigenschaften und Stimmungen der User zu gewinnen. Hierfür wurden 5 verschiedene Hypothesen entwickelt, die exemplarisch verschiedene Möglichkeiten aufzeigen, Informationen über die Kunden aus dem User Feedback zu extrahieren. Diese Analysen setzen die Hypothese voraus, dass geschriebene Texte Informationen über die Persönlichkeiten und Stimmungen der Verfasser offenbaren können. Diese Studie kann als Roadmap verwendet werden, um Informationen über die Persönlichkeiten und Stimmungen durch Analyse der App Store Reviews zu generieren.

Es wurden einige Studien durchgeführt, die davon ausgehen, dass die Art und Weise wie sich eine Person ausdrückt und deren Wortwahl Aufschluss über die Persönlichkeit und Stimmungslage dieser Person gibt. Durch Worte drückt man seine Gedanken und Gefühle aus und so verrät die Art zu kommunizieren viel über jemandes Charakter. Die moderne Psychologie und Psychotherapie geht davon aus, dass Erzählungen von Patienten deren unterbewusste Probleme reflektieren (z.B. Freud). Wortwahl und Redemuster können heutzutage mit Computer basierten Techniken analysiert und somit auch große Datenmengen auf einmal untersucht werden. Moderne Textanalyse-Methoden sind beispielsweise schon dazu in der Lage an Hand von Texten einer Person zu erkennen, ob diese lügt oder die Wahrheit sagt und welches Geschlecht die Person hat [BHS13] [TP10].

1.2. Motivation

Nun werden die Problemstellung und die Motivation vorgestellt, die zu der Durchführung dieser Arbeit geführt haben. In der App Entwicklung liegt häufig der Fall vor, dass nicht in Folge eines Auftrags entwickelt wird, sondern direkt für den freien Markt. Die Kunden und deren Bedürfnisse sind in einem solchen Fall nicht bekannt. Ist der Kunde beispielsweise ein Unternehmen, bekommt der Auftragsnehmer ein Lastenheft zur Verfügung gestellt, das die Anforderungen des Auftraggebers genau spezifiziert. Da sich diese im Laufe des Softwareentwicklungsprozesses meist ändern oder erst herausstellen, kann der Auftraggeber regelmäßiges Feedback geben. Diese Gegebenheiten sind bei der Entwicklung von Applikationen direkt für den Endbenutzer nicht gegeben.

Das User Feedback, das die Kunden zu den einzelnen Applikationen verfassen, hat daher einen großen betriebswirtschaftlichen Wert für die Hersteller. Es besitzt das Potential Informationen über die Kunden und deren Bedürfnisse zu liefern. Dieses Wissen ermöglicht eine Personalisierung der Applikationen, kann mögliche Probleme offenbaren und Verbesserungspotentiale aufdecken. Da die Datenmengen, die durch Reviews in App Stores entstehen, sehr umfangreich sind, kann das User Feedback nur durch sehr großen Aufwand manuell untersucht werden. Deswegen wird in dieser Studie ein Ansatz vorgestellt, um relevante Informationen über die Verfasser von App Store Reviews mit Hilfe von Data Mining Techniken und statischen Analysen zu generieren.

Die Analyse des App Store Feedbacks zur Generierung zusätzlicher Informationen über die Reviewer könnte an verschiedenen Stellen des Softwareentwicklungsprozesses eingesetzt werden. Abbildung 1.2 zeigt das Modell eines typischen Softwareentwicklungsprozesses. Ein Anwendungsfall wäre die Analyse der Reviews von ähnlichen Applikationen (Substituten) bei der Neuentwicklung oder aktuellen Versionen der Applikation bei der Software-Evolution während dem Requirements Engineering. 75 % der Wartungskosten dienen der Anpassung an sich ändernde Anforderungen oder Umgebungen. Abbildung 1.1 gibt einen Überblick über die Kernaufgaben des Requirements Engineering. Bei der Elicitation (Erhebung) der Anforderungen werden die Stakeholder der Software erfasst und deren Ziele definiert. An Hand dieser Ziele können die Anforderungen dann erhoben werden. Ein sehr wichtiger Stakeholder ist vor allem der User der App. Je mehr man über den User und dessen Charakteristiken weiß, desto besser kann die Anforderungserhebung sein und desto besser das Endergebnis. Denn im Endeffekt baut das System und dessen Features und Design auf diesen Anforderungen auf [BM13] [MB12].

Da man die Kunden und deren Alltag und Charakterzüge nicht kennt, ist es schwer die Anforderungen an neue Produkte oder Änderungen zu identifizieren. Eine Möglichkeit mit diesem Problem umzugehen sind Personas. Personas werden in der Softwareentwicklung, aber auch im Marketing eingesetzt, damit man ein besseres Verständnis über die tatsächlichen User oder die angestrebte Zielgruppe gewinnt. Personas sind fiktive Personen, die folgende Attribute besitzen können: Name, Alter, Beschäftigung, Familie, Freunde, Kleidungsstil, Vorlieben, sozioökonomischen Status, Ziele, Aufgaben,

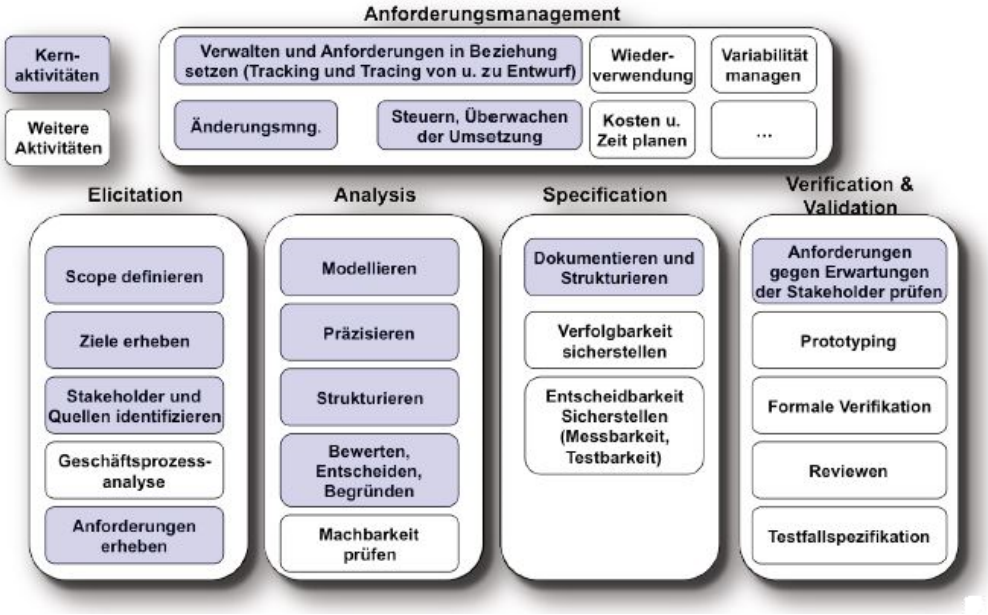


Figure 1.1.: Die wichtigsten Aktivitäten des Requirements Engineerings [BM13]

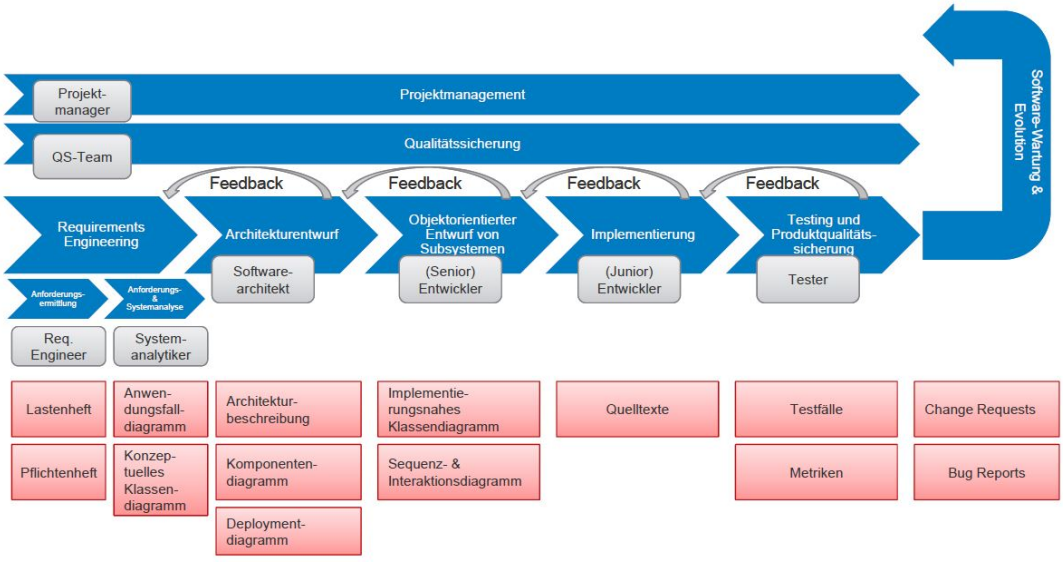


Figure 1.2.: Software Engineering Aktivitäten, Rollen und Artefakte im Überblick [MB12]

Lebensgeschichten usw. Zu einem Produkt werden in der Regel mehrere Personas erstellt, die dann repräsentativ dem Großteil der Kunden entsprechen. Auch ganze Marken können mit Personas ausgedrückt werden [GP].

Es wird gezeigt, dass die Analyse des User Feedbacks im App Store tatsächliche Informationen über die Persönlichkeiten der User liefert, welche verwendet werden können, um die Personas realistischer zu gestalten. Kennt man die Persönlichkeitseigenschaften der User, kann man diese den Personas zuweisen. Außerdem können mit Hilfe der Eigenschaften Vermutungen über den Lebensstil der User angestellt werden. Menschen, die extrovertierter und freundlicher sind, könnten beispielsweise ein ausgeprägteres soziales Leben haben, als die introvertierten, unfreundlichen User. Letztendlich kann damit das Requirements Engineering und somit auch das Endprodukt verbessert werden.

Eine weitere Einsatzmöglichkeit der Ergebnisse der User-Feedback-Analyse ist das Testing im Softwareentwicklungsprozess (Siehe Abbildung 1.2)während dem Beta-Testing (auch Akzeptanz- oder Abnahmetest). Beim Beta-Testing wird das Produkt erstmals entweder durch firmeninterne Mitarbeiter oder externe Beta-Tester (auch oft der Kunde) getestet. Vor allem durch externe Tester geht ein gewisses Risiko aus, da diese die fehlerhafte Software weiter geben könnten, wodurch es zu schlechter Publicity kommen könnte. Nichtsdestotrotz bietet zum Beispiel der Google Play Store die Möglichkeit, Apps im Play Store für Beta-Tester (selbst ausgewählte Nutzergruppen) frei zu schalten. Die Beta-Version der Applikation kann daraufhin nur von den Tester gesehen und herunter geladen werden. Die Tester geben ihr Feedback jedoch nicht im App Store ab, sondern übermitteln dies über einen, durch den Entwickler festgelegten, Kanal (z.B. Email) an den Entwickler. In dieser Phase des Softwareentwicklungsprozesses könnte es besonders hilfreich sein einen Überblick über die Stimmungen zu gewinnen, die mit der eigenen App assoziiert werden [DM93].

In dieser Studie wird gezeigt werden, dass die im Feedback enthaltenen Stimmungen nicht auf Grund von exogenen Einflüssen entstanden sind, sondern mit der App zusammen hängen. Die Sentiments können daher verwendet werden, um die Zufriedenheit der User mit der Applikation zu messen. Diese könnte auch auf Grund der Sterne-Bewertungen bestimmt werden, jedoch können mit den intervallskalierten Stimmungswerten genauere Ergebnisse als mit den ordinalskalierten Sterne-Bewertungen berechnet werden. Die Stimmungswerte sind den Sterne-Bewertungen also auf Grund der feineren Granularität vorzuziehen.

Die folgenden Seiten werden auf diese Weise strukturiert sein: Abschnitt 2 wird die Forschungsansätze ähnlicher Arbeiten beschreiben. Abschnitt 3 befasst sich mit den Hypothesen und den daraus folgenden Forschungsfragen, die im Zuge dieser Studie analysiert werden. Abschnitt 4 erklärt den Forschungsansatz und im darauffolgenden Abschnitt 5 werden die Ergebnisse der Analysen dargestellt. Abschnitt 6 beschreibt mögliche Interpretationen der Ergebnisse, sowie Ansätze für zukünftige Arbeiten. Im letzten Abschnitt 7 wird eine Schlussfolgerung der Erkenntnisse gezogen.

2. Related Work

Auswahl folgender Literatur:

[DC13]
[BHS13]
[**MichaelW.Macy2011**]
[**Rigby2007**]
[TP10]
[GRT11]
[**Chen2014**]

- On the Quest of Discovering Cultural Trails in Social Media?, Ruth Garcia
- Content-Based Similarity Measures of Weblog Authors, ChristopherWienberg, Melissa Roemmele, and Andrew S. Gordon

3. Hypothesen

Diese Studie soll zeigen, dass es möglich ist den App Store als Kanal zu verwenden, um Informationen über die User verschiedener Applikationen zu gewinnen. Exemplarisch wird ein Ansatz gezeigt, um das zu bewerkstelligen. Insbesondere wird in dieser Studie gezeigt, wie das User Feedback Informationen über die Persönlichkeitseigenschaften und Stimmungen der User offenbaren kann. Hierfür wurden fünf Hypothesen entwickelt, die Abhängigkeiten und Korrelationen zwischen den Eigenschaften und Stimmungen der User mit verschiedenen Faktoren vermuten. Um die Hypothesen zu bestätigen oder abzulehnen wurde für jede Hypothese exemplarisch eine entsprechende Forschungsfrage entwickelt. Die ersten drei Hypothesen betreffen die Persönlichkeitseigenschaften der User und die letzten zwei deren Stimmung. Die Hypothesen und die dazugehörigen Forschungsfragen sollen nun vorgestellt werden.

Die Analyse der Forschungsfrage **„Unterscheiden sich die Persönlichkeiten der Reviewer abhängig von der Applikation?“** ermittelt die Unterschiede der Persönlichkeitseigenschaften der Reviewer abhängig von der kommentierten Applikation. Die zu Grunde liegende Hypothese besagt, dass sich die Persönlichkeiten der User zwischen den einzelnen Applikationen unterscheiden. Es wird davon ausgegangen, dass Menschen mit ähnlichen Eigenschaften die selben Vorlieben besitzen, auch bei der Verwendung von Apps. Je unterschiedlicher die Applikationen, desto unterschiedlicher sind vermutlich auch die Charaktere der User. Diese Information ist von großer Bedeutung. Unterscheiden sich die Persönlichkeiten der User abhängig von der App, dann wäre es für den Softwareentwicklungsprozess wichtig die Persönlichkeiten der User zu kennen und die Applikationen darauf anzupassen. Denn je nachdem, wie die Eigenschaften eines Menschen ausfallen, unterscheiden sich auch dessen Erwartungen und Anforderungen an ein Produkt. Gibt es jedoch keine applikationsabhängigen Unterschiede in den Persönlichkeitswerten, spielen die Personas der User keine Rolle.

Bei der Forschungsfrage **"Unterscheiden sich die Persönlichkeiten der Reviewer abhängig von deren Herkunft?"** sollen regionale Unterschiede der Persönlichkeitseigenschaften überprüft werden. Die zu Grunde liegende Hypothese vermutet, dass sich die User der verschiedenen Märkte in ihren Persönlichkeitseigenschaften unterscheiden. Da Apple keine exakten Informationen über den Herkunftsort der Reviewer zur Verfügung stellt, kann man nur auf Grund der verwendeten App Stores auf deren Herkunft schließen. Die meisten Länder besitzen ihren eigenen App Store, aber manchmal werden auch mehrere kleine Länder (die geographisch nahe an einander liegen) in einem App Store zusammen gefasst. Hierbei gehen wir vereinfacht davon aus, dass der Reviewer auch aus der Region stammt, in dessen App Store er seinen Kommentar verfasst hat. Auch dieser Aspekt ist von großer Wichtigkeit. Bei international positionierten Produk-

ten (was die ausgewählten Applikationen auch sind) ist es manchmal notwendig, sie an die Kunden der einzelnen Nationen anzupassen. Bei der Internationalisierung eines Produkts kann nämlich manchmal das Problem auftreten, dass neue Märkte ein, in alten Märkten etabliertes, Produkt nicht akzeptieren, weil deren User andere Anforderungen besitzen. Für diese Zwecke müssen sich Anbieter mit nationalen Unterschieden ihrer Kunden auseinander setzen. Ob Unterschiede der Persönlichkeitseigenschaften der User zwischen verschiedenen Märkten allgemein und innerhalb der Applikationen existieren, wird bei Forschungsfrage zwei an Hand von fünf ausgewählten App Stores ermittelt.

Die dritte Forschungsfrage lautet:**„Unterscheiden sich die Persönlichkeiten der Reviewer, die eine positive, neutrale oder negative Sterne-Bewertung abgegeben haben?“**. Jeder User muss vor der Verfassung eines Kommentars ein Feedback in Form von Sternen vergeben. Mit Hilfe dieser Bewertung gibt der User seine Zufriedenheit mit der Applikation an. Die schlechteste Bewertung entspricht einem Stern und die beste Bewertung fünf Sternen. Ein 1 – 2 Sterne Rating wurde als negativ, 3 Sterne neutral und 4 – 5 Sterne als positiv substituiert. Diese Forschungsfrage ermittelt also, ob sich die Persönlichkeiten der unzufriedenen, neutralen und zufriedenen User von einander unterscheiden. Die Hypothese zu dieser Forschungsfrage geht davon aus, dass zwischen diesen beiden Aspekten eine Abhängigkeit besteht. Es wird vermutet, dass die Persönlichkeitseigenschaften der User Einfluss auf deren Bewertungsverhalten haben, also ein bestimmter Typ Mensch eher dazu neigt positive, negative oder neutrale Bewertungen abzugeben, als ein anderer. Für den Softwareentwicklungsprozess ist es relevant zu wissen, wie groß die Zufriedenheit der User mit der Applikation ist. Denn ein sehr zufriedener User wird die App häufiger verwenden, wodurch die Werbung, mit deren Hilfe bei kostenlosen Applikationen ein wirtschaftlicher Nutzen entsteht, häufiger gesehen wird. Sehr zufriedene User kaufen auch mit einer höheren Wahrscheinlichkeit kostenpflichtige Versionen der App oder Add-ons. Trifft die Hypothese zu, dass das Bewertungsverhalten durch die Persönlichkeiten der User beeinflusst wird, kann eine Korrektur der Sterne-Bewertungen in Betracht gezogen werden, um deren Aussagekraft zu vergrößern. Das heißt es könnte eine Normalisierung durchgeführt werden, bei der die Sterne-Bewertungen bestimmter User (mit bestimmten Persönlichkeitseigenschaften) verringert oder erhöht werden.

Mit Hilfe der Forschungsfrage vier **"Korrelieren die abgegebenen Sterne-Bewertungen mit den Sentiments der Reviewer?"** soll festgestellt werden, ob die Sentiments der User als deren Zufriedenheit mit der Applikation interpretiert werden kann. Die Sterne-Bewertung entspricht per Definition der Zufriedenheit der User. Bei den Sentiments könnte es jedoch sein, dass die Stimmung des Users auf Grund von exogenen Faktoren, zum Beispiel Uhrzeit oder Jahreszeit, entsteht. In dieser Hypothese wird jedoch davon ausgegangen, dass die Stimmung, die im User Feedback ausgedrückt wird, mit der Zufriedenheit mit der Applikation zusammenhängt. Kann dies tatsächlich festgestellt werden, könnte die Stimmung in weiteren Analysen als Variable zur Messung der Zufriedenheit verwendet werden. Dies wäre von Vorteil, da die Stimmung des Users eine feinere Granularität als die Sterne-Bewertung zur Verfügung stellt (intervallskalierte

Stimmung anstatt ordinalskalierte Sterne-Bewertung).

Die fünfte Forschungsfrage **"Unterscheiden sich die Sentiments der Reviewer innerhalb einer Applikation abhängig von der Versionsnummer?"** soll eine mögliche Anwendung der Erkenntnisse von Forschungsfrage vier aufzeigen. Sie soll ein Beispiel darstellen, wie die Sentiment-Analyse der Userkommentare während dem Betatesting von Nutzen sein könnte. Die zu Grunde liegende Hypothese besagt, dass die Sentiments der User abhängig vom Release einer App sind. Also zum Beispiel, dass die User bei einer Version überwiegend positiv kommentieren und nach Veröffentlichung eines neuen Releases unzufriedener sind. Solche Stimmungsänderungen könnten auf die dargestellte Weise gemessen werden und daraufhin den Softwareentwicklungsprozess beeinflussen. Denn so könnten die Hersteller einer App erfahren, wenn es Probleme mit neuen Versionen einer App gibt oder ob die User mit dem neuen Release zufriedener sind als mit dem alten.

Im folgenden Forschungsansatz wird beschrieben, wie vorgegangen wurde, um Antworten für die fünf Forschungsfragen zu generieren. Dieser Ansatz kann auch auf ähnliche Forschungsfragen mit verschiedenen Variablen angewendet werden.

4. Forschungsansatz

In diesem Kapitel wird der Forschungsansatz erläutert, der angewendet wurde, um Antworten auf die Forschungsfragen zu generieren. Abbildung 4.1 visualisiert die Vorgehensweise, auf die in diesem Abschnitt näher eingegangen werden soll.

Als erstes werden die Daten aus dem App Store geladen und in einer lokalen MySQL-Datenbank abgespeichert. Daraufhin werden die Kommentare und deren Titel unabhängig von den anderen Daten der Reviews extrahiert und aufbereitet, um die Ergebnisse der Text Analyse zu verbessern. Dieses „bereinigte“ Text Feedback kann daraufhin durch Anwendung des Tools LIWC auf Stimmungen und Persönlichkeitsdimensionen untersucht werden. Linguistic Inquiry and Word Count (LIWC) ordnet die Wörter eines Textes verschiedenen psychologischen und sprachlichen Dimensionen, wie zum Beispiel "Artikel" und "Positive Wörter", zu. Für jede Dimension sind Dictionaries definiert, also Listen von Wörtern, die die Zugehörigkeit der Wörter zu verschiedenen Dimensionen festlegen (ein Wort kann auch zu mehreren Dimensionen gehören). Letztendlich erhält der User für jede Dimension einen Wert, der angibt wie hoch der prozentuale Anteil der Wörter, die zu dieser Dimension zugeordnet werden konnten, vom Gesamttext ist. Die LIWC Ausgabe wird daraufhin dazu verwendet die Big Five Personality Traits und die Stimmungen der User zu berechnen. Die Big Five sind fünf verschiedene Werte für Persönlichkeitseigenschaften, die mehreren Studien zu Folge den Charakter eines Menschen am besten beschreiben. Alle gewünschten Metadaten zu den Texten, wie z.B. Versionsnummer, Applikationsname und Datum, werden gemeinsam aus der Datenbank extrahiert und daraufhin für das Tool R [1] angepasst. Letztendlich können diese transformierten Metadaten, die Big Five Werte und Stimmungswerte, sowie weiterhin relevante Dimensionen des LIWC Outputs zusammengeführt werden. Auf diese Daten werden anschließend statistische Tests angewendet, um die Forschungsfragen zu beantworten. Im Folgenden sollen die wichtigsten Schritte genauer erklärt werden.

4.1. App Store Daten

4.1.1. Generierung

Als erstes müssen die Daten für die Analysen ausgewählt und gesammelt werden. Für die Extrahierung der Reviews aus dem Apple App Store wird ein leicht abgeändertes Open Source Scraping Tool verwendet[2]. Mit Hilfe dieses Tools werden die benötigten

[1] http://www.r-project.org/

[2] https://github.com/oklahomaok/AppStoreReview

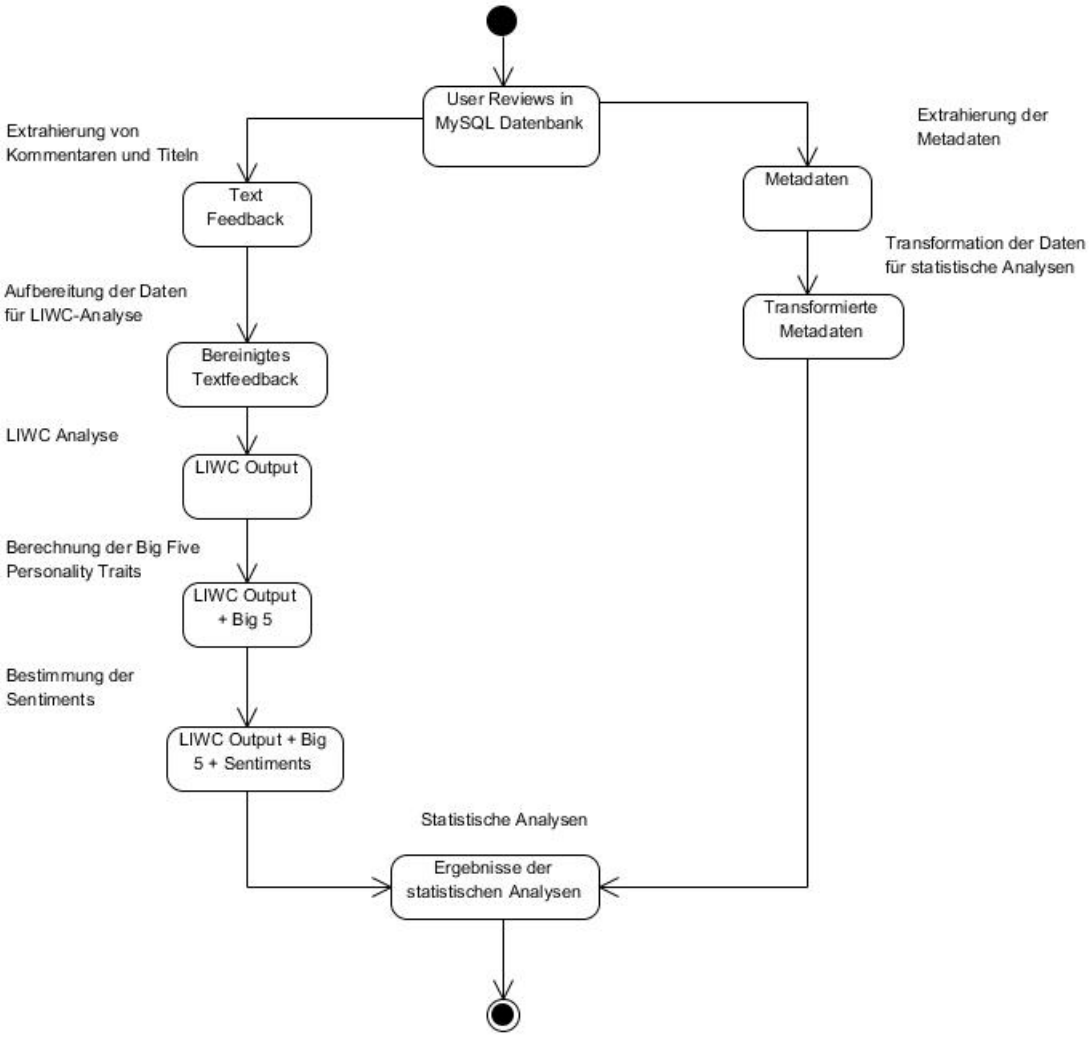


Figure 4.1.: Der Forschungsansatz im Überblick

Daten in eine lokale MySQL Datenbank geladen. Zur Erzeugung und Administration dieser Datenbank wird das Open Source Tool WampServer[3] benutzt, welches MySQL-Server und Web-Server zur Verfügung stellt. Außerdem bietet WampServer eine Weboberfläche an, mit der die Administration der Datenbank möglich ist (phpMyAdmin).

4.1.2. Überblick über den Datensatz

Die Daten, die im Zuge dieser Studie analysiert werden, stammen aus dem Apple App Store[4]. Sie umfassen den Namen, den der Reviewer angeben hat, das Datum an dem der Kommentar verfasst wurde, Titel und Kommentar des Reviews, die Anzahl der vergebenen Sterne und die Versionsnummer der kommentierten Applikation. Zusätzlich generierte Daten sind die ReviewID und die ProjectID, welche angibt zu welcher Applikation der Review angefertigt wurde. Außerdem geben zwei weitere Variablen Auskunft über den Namen der Region, dessen App Store verwendet wurde, um den Kommentar zu verfassen und eine Variable, die angibt ob eine negative, positive oder neutrale Sterne-Bewertung abgegeben worden ist.

Für die gewählten Forschungsfragen ist es notwendig Kommentare verschiedener App Stores, also unterschiedlicher Regionen, zu laden. Diese Regionen sollten möglichst unterschiedlich bezüglich ihrer Kultur sein, weshalb versucht wurde geeignete Regionen von verschiedenen Kontinenten auszuwählen. Es konnten nur Regionen in Betracht gezogen werden, welche eine rege App Store Aktivität vorweisen konnten, bzw. wo genug User-Kommentare für die Analysen vorhanden waren. In dieser Studie wird das englische Dictionary von LIWC verwendet, das heißt das Tool erkennt nur englische Wörter und kann nur diese den verschiedenen Dimensionen zuordnen. Daher mussten Regionen gewählt werden, in denen der Großteil der Reviews auch in Englisch verfasst wurde. Diesen Überlegungen zu Folge fiel die Wahl auf die App Stores von Singapur (Asien), den Vereinigten Staaten (Nordamerika), Kanada (Nordamerika), Großbritannien (Europa) und den App Store mehrerer Westafrikanischer Länder (Afrika). Unter dem Begriff „Afrika-Länder“ werden die afrikanischen Länder Elfenbeinküste (ci), Kamerun (cm), Zentralafrikanische Republik (cf), Guinea (gn), Äquatorialguinea (gq) und Marokko (ma)[5] zusammengefasst, da diese über einen gemeinsamen App Store verfügen. Dieser App Store wurde in die Analyse mit einbezogen, da man eine Land aus Afrika verwenden wollte, die meisten App Stores (für einzelne Länder) jedoch zu wenig Kommentare umfasst hatten [CC].

Bei der Auswahl der Applikationen musste darauf geachtet werden, internationale Apps zu wählen, welche in allen ausgesuchten Regionen ausreichend viele Kommentare im untersuchten Zeitraum (hier das Jahr 2013) besitzen. Um die Analysen möglichst vielfältig zu halten, wurden Applikationen unterschiedlicher Kategorien und Größen (Anzahl Userkommentare) verwendet. Ausgewählt wurden letztendlich die User-

[3] http://www.wampserver.com/en/
[4] https://itunes.apple.com/us/genre/ios/id36
[5] In Klammern die Country Codes der Länder im Apple App Store

Kommentare der acht Applikationen Angry Birds (Games), Dropbox (Productivity), Evernote (Productivity), Adobe Reader (Business), Tripadvisor (Travel), Booking.com (Travel), Pinterest (Social Networking), und WhatsApp Messenger (Social Networking)[6].

Die ausgewählten Applikationen sollen nun kurz vorgestellt werden, da ihre Verwendung relevant für die Interpretation der Ergebnisse ist. Angry Birds ist ein Spiel von großer Popularität. Das Ziel ist es Schweine, die Vögeln ihre Eier geklaut haben, mit diesen abzuwerfen. Dropbox bietet 2 GB Speicherplatz in einer Cloud an, sodass Daten gespeichert und auf verschiedenen Endgeräten zugegriffen werden können. Evernote wird dazu verwendet Notizen und mit diesen Notizen verbundene Links, Fotos und Sprachmemos zu verwalten. Mit Hilfe von Adobe Reader ist das Öffnen und die Verwaltung von PDF-Dateien möglich. Die Applikation Tripadvisor unterstützt die Urlaubsplanung, indem sie Bewertungen von Hotels, Restaurants und Aktivitäten in verschiedenen Ländern zur Verfügung stellt und die Buchung von diesen unterstützt. Die Booking App vereinfacht das Suchen und Finden von geeigneten Hotels in verschiedenen Städten. Auch hier werden Bewertungen von ehemaligen Kunden angezeigt. Die App Pinterest kann dazu verwendet werden Ideen, Bilder, Videos und weitere Daten in "Pins" zu organisieren und im Rahmen einer Social Media Umgebung zu "posten". WhatsApp bietet die kostenlose Möglichkeit Textnachrichten über das Internet auszutauschen.

Da die Applikationen unterschiedliche lange existieren und zu unterschiedlichen Zeiten in den verschiedenen Ländern eingeführt wurden, wurden nur die Reviews aus dem Jahr 2013 für die statistischen Analysen verwendet. Außerdem ist es notwendig, zu kurze Kommentare aus den Analysen auszuschließen. Userkommentare in App Stores sind oft sehr kurz, was eine schlechte Verteilung der Daten zu Folge haben kann. Zum Beispiel ein kurzer Kommentar wie "I love this app!" würde nur in den Dimensionen "First-Person-Singular", "Social", "Present tense" und "Positive Emotion" zu einem Wert ungleich Null führen. Auf diese Weise entstehen viele Nullwerte in den anderen Big Five Personality Traits Dimensionen, die bewirken, dass die Daten sehr weit von einer Normalverteilung abweichen. Das ist problematisch, weil normal verteilte Daten für viele statistische Tests Voraussetzung sind. Um diesen Effekt einzudämmen, werden ausschließlich Reviews, die eine Länge von 20 Wörtern überschreiten, für die Analysen eingesetzt. Außerdem wird durch diese Vorgehensweise die Aussagekraft der Big Five erhöht, da sehr kurze Texte zu wenig Informationen bieten, um Aussagen über die Persönlichkeit des Verfassers zu treffen.

Von den insgesamt 2 236 456 geladenen Kommentaren werden daher nur 119 690 Stück in den Analysen berücksichtigt. Ein Überblick über die geladenen Daten und denen, die letztendlich für die Analyse verwendet wurden, ist in Tabelle 4.1 zu sehen.

[6]In Klammern die App Kategorien des Apple App Stores

Table 4.1.: Überblick über Daten im App Store

	U.S.A.	Afrika-Länder	Großbritannien	Kanada	Singapur	Insgesamt
Alle Kommentare	898,255	1,047,508	206,239	65,566	18,888	2,236,456
Jahr 2013	174,160	173,264	24,031	16,499	3,341	391,295
Jahr 2013 und > 20 Wörter	52,745	52,383	8,345	5,088	1,129	119,690

Table 4.2.: Überblick über Daten, die in der Analyse verwendet wurden (Jahr 2013 und > 20 Wörter)

	U.S.A.	Afrika-Länder	Großbritannien	Kanada	Singapur	Insgesamt
Angry Birds	704	703	132	704	1	2,244
Dropbox	1,203	1,202	257	138	19	2,819
Evernote	5,308	5,288	889	400	54	11,939
Adobe Reader	310	307	74	50	9	750
Tripadvisor	1,581	1,566	725	187	23	4,082
Booking	240	238	222	51	15	766
Pinterest	38,225	37,998	2,636	2,935	50	81,844
WhatsApp	5,174	5,081	3,410	623	958	15,246

4.2. Textdaten

4.2.1. Laden der Textdaten

Die Textdaten sind die Daten, die mittels des Tools LIWC auf sprachliche und psychologische Dimensionen untersucht wurden. Es ist es notwendig eine Verbindung zu der MySQL Datenbank herzustellen, woraufhin mittels SQL-Befehlen diese Textdaten aus der Datenbank geladen werden können. Bei der Abspeicherung der Daten muss berücksichtigt werden, dass das LIWC-Tool ganze Text-Files analysiert. Da in dieser Studie nur Titel und Kommentar des Reviews für die Textanalyse berücksichtigt werden sollen, müssen alle anderen Informationen, wie Reviewer und Versionsnummer (Metadaten), in ein separates File geschrieben werden. Andernfalls würden beispielsweise Reviewer-Namen wie „crazy guy“ auch in die Analyse mit einfließen.

Jeder Kommentar muss vereinzelt in ein Textfile geschrieben werden, zusammen mit dem Titel des Kommentars, um unabhängig von den anderen Daten analysiert werden zu können. Um die Analyse des Textfiles und die entsprechenden Metadaten später wieder zusammen zu führen, wird eine eindeutige ID benötigt. Da die ReviewID nur innerhalb eines Landes eindeutig ist, muss an die ReviewID noch das Kürzel des entsprechenden Landes angehängt werden.

4.2.2. Vorbereitung der Daten auf die Analyse

Dann wurden die Daten auf die LIWC Analyse vorbereitet. Im Operator's Manual [PBF], dass beim Kauf einer LIWC-Lizenz zur Verfügung gestellt wird, gibt es einen Abschnitt zum Thema „Preparing Written Text For LIWC2007 Analysis“. An dieser Stelle erklären die Entwickler des Tools, welche Veränderungen an den Daten (hier insbesondere den Textdaten) durchgeführt werden sollten, um Fehler in der Analyse zu vermeiden.

Entsprechend dieser Empfehlungen wurden zunächst einige Zeichen durch andere ersetzt. Beispielsweise schreibt Pennebaker [PBF], dass „'s“ als Possessivpronomen („Birgit's flowers“) registriert wird, was bei der Kurzform des Verbs „is“ („Birgit's cooking“) natürlich falsch wäre. Da es für dieses Problem keine gute Lösung gibt, ohne den Kontext in Betracht zu ziehen, werden alle „'s“ durch „is“ ersetzt, da die Kurzform des Verbs „is“ um einiges häufiger Verwendung findet als das Possessivpronomen [PBF].

Des Weiteren müssen HTML Codes, die sich noch im Text befunden hatten, durch deren decodierte Symbole ersetzt werden. Zum Beispiel entspricht der HTML-Code "'" dem Zeichen "'". Einige Sonderzeichen wie „@“ und „&“ werden mit deren textuelle Namen, also „at“ und „and“ ausgetauscht. Außerdem weist Pennebaker darauf hin, dass im Fall von mit Bindestrich verbundene Phrasen, wie „this-or-that“, LIWC nach einem Eigenwort im Dictionary suchen wird. Um dies zu verhindern, reicht es den Bindestrich durch Leerzeichen hervorzuheben. Dadurch können die einzelnen Wörter auf beiden Seiten des Bindestrichs gewertet werden. Da das aktuelle Englisch Dictionary keine Eigenwörter mit Bindestrich enthält, können alle „-“ durch „ - “ ersetzt werden.

Für die Korrektheit der LIWC Dimension „Words/Sentence“ müssen ebenfalls ein

paar Änderungen vorgenommen werden. Jedes abschließende Satzzeichen, wie ein Punkt oder Fragezeichen, wird als Ende des Satzes interpretiert. Insbesondere Punkte kommen jedoch oft innerhalb von Sätzen vor, zum Beispiel bei Abkürzungen wie „Jan." für „January" oder bei Versionsnummern (z.B Version Nummer 1.2.4), über die geschrieben wird. Um die Kategorie „Words/Sentence" aussagekräftig zu machen, wurden häufig gebrauchte Abkürzungen durch deren ausgeschriebene Form ersetzt und die Punkte der Versionsnummern entfernt. Aus dem selben Grund wurden Punkte nach Kommentaren und Titeln eingefügt, wenn diese vom Verfasser weggelassen worden waren.

4.3. Metadaten

Die Metadaten müssen gesondert von Kommentar und Titel aus der Datenbank geladen und gespeichert werden, da diese sonst die LIWC-Analyse beeinflussen würden. Mit Metadaten sind folgende Informationen der Reviews gemeint: ReviewID, die Anzahl der vergebenen Sterne, der Reviewername, das Verfassungsdatum des Kommentars und die Versionsnummer zum Zeitpunkt der Bewertung. Die ReviewID ist nur innerhalb eines Landes eindeutig (wegen der Organisation der Daten in der Datenbank) und wird deswegen noch mit einem Kürzel für das entsprechende Land versehen. Zusätzlich mit Python-Scripts generierte Metadaten sind Applikationsname und Land des entsprechenden App Stores. Außerdem wird eine Variable erzeugt, die angeben soll, ob die Vergabe der Sterne auf eine positive (4-5 Sterne), neutrale (3 Sterne) oder negative (1-2 Sterne) Bewertung schließen lässt. Die letzte Variable wird für Forschungsfrage drei benötigt.

Da all diese Informationen durch Anwendung des Statistik Tools R verarbeitet werden sollen, ist es notwendig die Daten an die Limitationen des Tools anzupassen. Bei der Erstellung der Tabelle in R über die oben genannten Metadaten, traten Probleme bei Leerzeichen oder manchen Satzzeichen wie "'" auf. Diese Zeichen wurden deswegen nach Extrahierung der Metadaten aus diesen entfernt. Diese Schritte wurden ebenfalls Hilfe von Python-Scripts durchgeführt.

4.4. LIWC – Analyse der Textdaten

4.4.1. Linguistic Inquiry and Word Count Tool (LIWC)

Für die psychologische Analyse der Daten wurde das kommerzielle Tool „Linguistic Inquiry and Word Count (LIWC)" von Pennebaker Conglomerates Inc. angewendet. LIWC ist eine Software zur Textanalyse von psychologischen Strukturen durch den Abgleich von Wörtern mit verschiedenen Wortkategorien und soll nun genauer vorgestellt werden [Pen+].

Die Besonderheit des Tools sind die Dictionaries in unterschiedlichen Sprachen, deren Wörter 82 unterschiedlichen Wortkategorien, auch „Dimensionen" genannt, zugewiesen wurden. Lässt man einen Text mittels LIWC analysieren, werden die Wörter des Textes

(target words) Wort für Wort mit dem Dictionary (dictionary words) der gewünschten Sprache und Version abgeglichen. Hierbei gehört ein Wort meist zu mehr als einer Dimension. Beispielsweise wurde das Wort „look" den Dimensionen "verb", "percept", "present" und "see" zugewiesen. Wird ein target word im Dictionary gefunden, ist darin auch vermerkt welchen Kategorien es angehört, welche daraufhin inkrementiert werden. Insgesamt ist das Ziel, dass für einen Text der prozentuale Anteil der einzelnen Dimensionen berechnet wird. Die Dimensionen und ihre prozentualen Anteile werden als Tabelle angeordnet in einem Text-File gespeichert. Analysiert man mehrere Text-Files gleichzeitig wird jedes File als Reihe einer Tabelle aufgelistet [Pen+].

Die LIWC-Dimensionen sind vielseitig und decken unterschiedliche Bereiche ab. Welche Dimensionen im Zuge der Analyse berechnet und deren Anteile ins Output-File geschrieben werden sollen, kann vor Beginn der Analyse eingestellt werden [Pen+].

Einerseits gibt es „Linguistic Processes“ Dimensionen, also Dimensionen sprachlicher Prozesse, die weniger auf die Wortwahl eines Menschen abzielen, sondern auf die Art und Weise wie sich dieser ausdrückt. Hier wird beispielsweise überprüft, ob ein Wort ein Verb (Common verbs, Auxiliary verbs) oder Artikel (Articles) ist. Ob dieses Wort als Schimpfwort (swear words) bezeichnet werden kann und um welche Art von Pronomen (1st pers singular, 1st pers plural usw.) es sich handelt. Außerdem werden Eckdaten zu dem Text berechnet, wie zum Beispiel die Anzahl der Wörter, die eine Übereinstimmung mit dem Dictionary aufgewiesen haben (Dictionary words) und die insgesamt in einem Text verwendet wurden (Word Count) [LIWC].

Zu beachten ist, dass innerhalb der Dimensionen übergeordnete und untergeordnete Wortkategorien definiert wurden. Ein Pronomen beispielsweise wird zunächst den Zähler der Wortkategorien „Total function words“ und „Total pronouns“ erhöhen. Und je nachdem ob dieses Pronomen ein Personalpronomen ist oder nicht, wird nach Erfassung des Worts der Zähler weiterer Kategorien inkrementiert [LIWC].

Mit Hilfe der „Psychological Processes“ Dimensionen soll die Wortwahl eines Menschen Einblicke in dessen Psyche geben. Die prozentual häufige Verwendung von Wörtern wie „buddy“ (friends) und „boy“ (humans) könnte zum Beispiel darauf hinweisen, dass eine Person ein gesundes Sozialleben besitzt. Die Untergruppe „Relativity“ mit den Wortkategorien „Motion“, „Time“ und „Space“ könnte hingegen Informationen über die geistige Gesundheit eines Menschen offenbaren [LIWC].

Für psychologische Analysen sind jedoch nicht nur die „Psychological Processes“ von Bedeutung. Bei den Analysen dieser Studie wird vor allem von der Theorie ausgegangen, dass die Redeweise eines Menschen auch unterbewusste Persönlichkeitsstrukturen offenbaren kann und daher die „Linguistic Processes“ möglicherweise noch mehr über einen Menschen verraten können, als die „Psychological Processes“ Dimensionen.

Des Weiteren können Texte auch auf „Personal Concerns“ Dimensionen untersucht werden, mit Hilfe derer überprüft werden kann, ob viele Wörter aus den Bereichen „Leisure“, „Work“, „Religion“ usw. genutzt werden. Schließlich können die „Spoken categories“ Dimensionen im Falle einer Analyse von mündlichen Protokollen oder Social Media Texten „Fillers“ wie „I mean“ und „You know“ oder „Nonfluencies“ erken-

nen. "Nonfluencies" sind Wörter wie „hm“ und „umm“, die darauf schließen lassen, dass sich die sprechende Person nicht flüssig ausdrücken kann [LIWC]. Interessierte haben die Möglichkeit, auf der Homepage der Softwareanbieter von LIWC das Tool auszuprobieren[7].

4.4.2. Anwendung

Nach der Bereinigung des Text Feedbacks aus dem App Store ist dieses bereit für die LIWC Analyse. Zur Bestimmung der Persönlichkeitseigenschaften benötigt man die Dimensionen "First-Person-Singular", "Positive Emotion", "Negative Emotion", "Tentativeness", "Negations", "Social", "Articles", "Words of more than 6 letters" und "Present tense" von LIWC. Um zu kurze Kommentare zu identifizieren und aus der Analyse auszuschließen wird ebenfalls die Dimension "Word Count" benötigt. Diese Werte müssen für jeden einzelnen Kommentar generiert werden.

4.5. Bestimmung der Persönlichkeitseigenschaften

4.5.1. Big Five Personality Traits Ansatz

Bei den Forschungsfragen dieser Studie werden einige Forschungsfragen untersucht, deren Ziel es ist, Unterschiede von App Store Reviewern in deren Persönlichkeitseigenschaften zu identifizieren. Hierfür müssen messbare Persönlichkeitseigenschaften festgelegt werden, die im Zuge der Analyse gemessen und verglichen werden. Nun soll der Big Five Personality Traits Ansatz vorgestellt werden, der hierfür in dieser Studie verwendet wurde. Der Big Five Personality Traits Ansatz wird in der Literatur oft als die beste Möglichkeit bezeichnet, um die Persönlichkeit einer Person zu beschreiben. Diesem Ansatz zu Folge setzt sich die Persönlichkeit eines Menschen vor allem aus fünf Komponenten zusammen, wobei die Stärke der einzelnen Eigenschaften individuell ist: Openness, Conscientiousness, Extraversion, Agreeableness und Neuroticism [WRG13].

Dieses Modell wurde Mitte des 20. Jahrhunderts durch das Zusammenspiel mehrerer Forschungsarbeiten entwickelt. Raymond B. Cattell begann 1943 mit Faktorenanalysen über mehrere Persönlichkeit beschreibende Ausdrücke. Die 12 Faktoren, welche Cattell letztendlich identifizierte, wurden einige Jahre darauf von mehreren Forschern wie Digman und Takemoto-Chock weiter analysiert. Diese und weitere Forscher bestimmten schließlich fünf Persönlichkeitseigenschaften, deren Ausprägungen am besten die Persönlichkeit eines Menschen darstellen können [Gol+90].

Erstens die Offenheit für neue Erfahrungen (Openness), was zum Beispiel auch bedeuten kann, dass ein Mensch fantasievoll und neugierig ist oder eine künstlerische Veranlagung besitzt. Als zweite Komponente gibt es das Pflichtbewusstsein (Conscientiousness), das das Maß an Ausdauer, Verantwortungsbewusstsein, Verlässlichkeit und Ehrgeiz eines Menschen angibt. Ist jemand extrovertiert (Extraversion), die dritte

[7] http://liwc.net/tryonline.php

Komponente, dann lässt das darauf schließen, dass es sich um eine offene und selbstbewusste Person handelt, die gerne Umgang mit anderen Menschen pflegt. Die vierte Eigenschaft Freundlichkeit oder Umgänglichkeit (Agreeableness) beschreibt hingegen Menschen, die hilfsbereit und entgegenkommend sind. Menschen mit einem hohen Anteil an „Agreeableness“ sind friedliche, vertrauenswürdige Personen, die Konflikte lieber vermeiden möchten. Die letzte Big Five Eigenschaft gibt an wie neurotisch eine Person ist (Neuroticism), das heißt wie unsicher und ängstlich, also deren emotionale Stabilität. Neurotische Menschen sind launisch und haben eine negativere Grundeinstellung [GRT11].

Besitzt eine Person eine signifikant niedrige Menge einer der fünf Personality Traits, sagt das im umgekehrten Sinn genauso viel über deren Persönlichkeit aus. Beispielsweise ist ein Mensch mit einem niedrigen Grad an Neuroticism sehr selbstsicher und emotional gefestigt.

4.5.2. Berechnung der Big Five

Zur Berechnung der Big Five werden Ergebnisse einer Studie von J. Pennebaker und L. King [PK99] verwendet. Diese untersuchten unter anderem die Korrelationen zwischen den LIWC Dimensionen und den Big Five Personality Traits. Zur Gewinnung der Daten mussten 1 203 angehende Psychology Studenten im Rahmen eines Einführungskurses für Psychology über das Semester hinweg einige Fragen schriftlich beantworten (die Daten wurden in Verlauf von sieben Jahren gesammelt). Mit diesen Informationen wurden daraufhin die Big Five Scores, sowie die LIWC Dimensionen der Texte bestimmt. Deren Korrelationen sind in Tabelle 4.3 zu sehen. Für die Analysen dieser Studie werden ausschließlich die Korrelationen, die auf einem 0,01 Niveau signifikant sind berücksichtigt. Zur Berechnung der Big Five Scores kann man vereinfachend positive Korrelationen addieren und negative Korrelationen subtrahieren, sodass folgende Formeln entstehen:

$$Neuroticism = +First\text{-}Person\text{-}Singular + Negative\ Emotion - Positive\ Emotion$$
$$Extraversion = +Social + Positive\ Emotion - Tentativeness - Negations$$
$$Openness = +Articles + Words\ of\ more\ than\ 6\ letters + Tentativeness - First\text{-}Person\text{-}Singular - Present\ tense$$
$$Agreeableness = -Articles$$
$$Conscientiousness = -Negations - Negative\ Emotion$$

Diese Formeln werden nun eingesetzt, um aus dem LIWC Output, der in Abschnitt 4.4.2 generiert wurde, die <Werte für die fünf Persönlichkeitseigenschaften zu gewinnen. Hierfür werden die Daten in das Tool R eingelesen und dort verarbeitet. Die Eigenschaftswerte entstehen durch die Addition und Subtraktion verschiedener Prozentzahlen. Einige Wörter können gleichzeitig in mehreren Dimensionen enthalten sein. Bei den für die Big Five Personality Traits relevanten Dimensionen gibt es zwischen folgenden Dimensionen gemeinsame Wörter: "Social" & "Positive Emotion" (z.B. "encourage"),

LIWC factor	Neuroticism	Extraversion	Openness	Agreeableness	Conscien.
Immediacy	**.10***	**.04**	**-.06****	**.07***	**-.02**
First-person singular	.13**	.04	-.13**	.07*	.01
Articles	-.09*	-.09*	-.13**	-.15**	-.04
Words of more than 6 letters	-.03	-.04	.16**	-.03	-.06
Present tense	.06	.01	-.15**	.04	.00
Discrepancies	.05	-.03	-.01	-.02	-.07*
Making Distinctions	**.05**	**-.14****	**.06**	**-.05**	**-.13****
Exclusive	.00	-.08*	.10*	-.06	-.08*
Tentativeness	.06	-.14**	.11**	-.02	-.06
Negations	.05	-.12**	.00	-.04	-.15**
Inclusive	-.01	.07*	.01	.03	.06
The Social Past	**.04**	**.00**	**.08***	**-.02**	**-.04**
Past tense	.03	.04	-.03	.06	-.06
Social	-.01	.12**	.02	.00	.02
Positive emotion	-.13**	.15**	-.06	.07*	.07*
Rationalization	**-.06**	**.02**	**-.03**	**.07**	**.04**
Insight	.03	-.02	.07*	.05	-.01
Causation	.03	-.08*	-.08*	.00	-.07*
Negative Emotion	.16**	-.08*	.05	-.07*	-.15**

Table 4.3.: Korrelation der LIWC Dimensionen mit den Big Five Personality Dimensions. * $p < 0.05$, ** $p < 0.01$, two-tailed [PK99].

"Words of more than 6 letters" & "Tentativeness" (z.B. "guessing"), "Articles" & "Tentativeness" ("a lot"), "First-Person-Singular" & "Present tense" (z.B. "I'm") Die Big Five Personality Traits können daher in folgenden Intervallen liegen:

Neuroticism $\in [-100, 100]$
Extraversion $\in [-100, 200]$
Openness $\in [-200, 200]$
Agreeableness $\in [-100, 0]$
Conscientiousness $\in [-100, 0]$

Außerdem müssen vor Berechnung der statistischen Analysen die Big Five Scores und alle anderen relevanten Daten (z.B. die Metadaten) wieder zusammengeführt werden.

4.6. Berechnung der Sentiments

Neben den Persönlichkeitseigenschaften der User wird in dieser Studie auch die Stimmung der User berechnet und analysiert. Hierfür wurden die Dimensionen "Positive Emotion" und "Negative Emotion" verwendet, also die Prozentsätze der negativen oder positiven Wörter vom Gesamttext. Jedem Kommentar wurde entweder der mit (-1) multiplizierte Wert für "Negative Emotion" oder der "Positive Emotion" Wert zugewiesen. Da negative Emotionen weniger stark ausgedrückt werden als positive Emotionen wurde bei einem Review, bei dem "Negative Emotion" $* 1,5 >=$ "Positive Emotion" der Wert $(-1) *$ "Negative Emotion", andernfalls der "Positive Emotion" Wert festgesetzt (Gewichtung der negativen Emotionen). Diese Werte werden dann als die Stimmung oder die Sentiments des Kommentars bezeichnet [SentiStr].

Da die Werte für "Negative Emotion" und "Positive Emotion" Prozentzahlen sind und "Negative Emotion" mit (-1) multipliziert wird, befinden sich die Sentiments immer in dem Intervall $[-100, 100]$ (besitzen auch keine gemeinsamen Wörter).

4.7. Statistische Analysen

In dieser Section werden die statistischen Methoden beschrieben, die eingesetzt wurden, um Antworten auf die Forschungsfragen zu generieren. Hierbei wird genauer darauf eingegangen, warum die einzelnen Methoden ausgewählt wurden und was mit diesen berechnet werden kann.

4.7.1. Vortests

Vor der Auswahl geeigneter Methoden muss man erst Informationen über den eigenen Datensatz gewinnen. Da jeder Test spezifische Voraussetzungen besitzt, ist es notwendig zu wissen, welche Annahmen die Daten erfüllen oder verletzen.

Für die Forschungsfragen, die in Kapitel 3 beschrieben werden, benötigt man geeignete Varianzanalysen, Post-Hoc-Tests und Korrelationstests. Innerhalb dieser Verfahren gibt es parametrische und nichtparametrische (oder auch verteilungsunabhängige) Tests. Parametrische Tests haben eine größere Teststärke (Power) und wären dadurch zu bevorzugen. Um parametrische Tests anwenden zu können, müssen die Daten jedoch mehrere Bedingungen erfüllen. Bei zu starker Verletzung der notwendigen Voraussetzungen, müssen nichtparametrische Tests angewendet werden, wenn man verlässliche Ergebnisse gewinnen möchte. Nichtparametrische Tests werden auch Rangsummentests genannt. Die statistische Entscheidung mit Hilfe eines Rangsummentests ist eher konservativ, das heißt er hält länger an der Nullhypothese fest, als nötig wäre. Große Stichproben führen hingegen zu einer Verbesserung der Power der meisten Tests und wirken diesem Effekt entgegen [JB87].

Normalverteilung

Eine häufige Voraussetzung an die Daten ist, dass sie normalverteilt sind. Insbesondere für die parametrischen Varianzanalysen und Post-Hoc-Tests wird gefordert, dass die abhängigen Variablen in jeder Gruppe einer Normalverteilung folgen müssen. Die Gruppen sind hierbei die unterschiedlichen Teilmengen der Grundgesamtheit, die durch die unabhängige Variable definiert werden. Ob es zwischen diesen Gruppen einen Unterschied gibt, soll mit Hilfe der Varianzanalysen und Post-Hoc-Tests festgestellt werden. Bei der ersten Forschungsfragen gibt es zum Beispiel acht Gruppen, jeweils eine für die Daten jeder Applikation. Es ist folglich notwendig für jede Forschungsfrage eigene Tests auf Normalverteilung durchzuführen, da diese auf jede, durch die unabhängige Variable definierte, Gruppe angewendet werden müssen.

Um einen ersten Überblick über den Datensatz zu gewinnen, eignen sich besonders gut Quantile-Quantile-Plots (QQ-Plots). Diese bieten eine Möglichkeit zu überprüfen, ob "zwei Messwertreihen aus Grundgesamtheiten mit der gleichen Verteilung stammen" [HS12]. Hierfür werden die Quantile der ersten Messwertreihe gegen die der zweiten im Koordinatensystem aufgetragen. Gehören die beiden Messwertreihen der gleichen Grundgesamtheit an, muss der resultierende Graph der Winkelhalbierenden (45° Linie) gleichen (mit nur kleinen Abweichungen). Möchte man überprüfen, ob ein Datensatz normalverteilt ist, kann man dessen Messwertreihe durch QQ-Plots gegen die Messwertreihe der Normalverteilung auftragen. Diese Methode wurde auf den Datensatz der Studie angewendet, was zu der Vermutung führte, dass keine Normalverteilung vorlag. Trotz der Limitierung des Datensatzes auf Kommentare mit mehr als 20 Wörtern waren bei der LIWC-Analyse einige Nullwerte entstanden (z.B. posemo=0.00, social=0.00). Diese Werte verzerrten als "Ausreißer" die Verteilung der Daten, weshalb es zu Abweichungen von der Normalverteilung gekommen war. In Abbildung 4.2 sind zur Verdeutlichung die QQ Plots der Big Five der Evernote Daten aus dem App Store der Afrika-Länder zu erkennen. Die restlichen QQ Plots befinden sich im Appendix. [HS12].

Um den Verdacht zu überprüfen, dass die Daten nicht normalverteilt sind, wurde in dieser Studie der Jarque-Bera-Test [JB87] verwendet. Der Jarque-Bera-Test ist ein

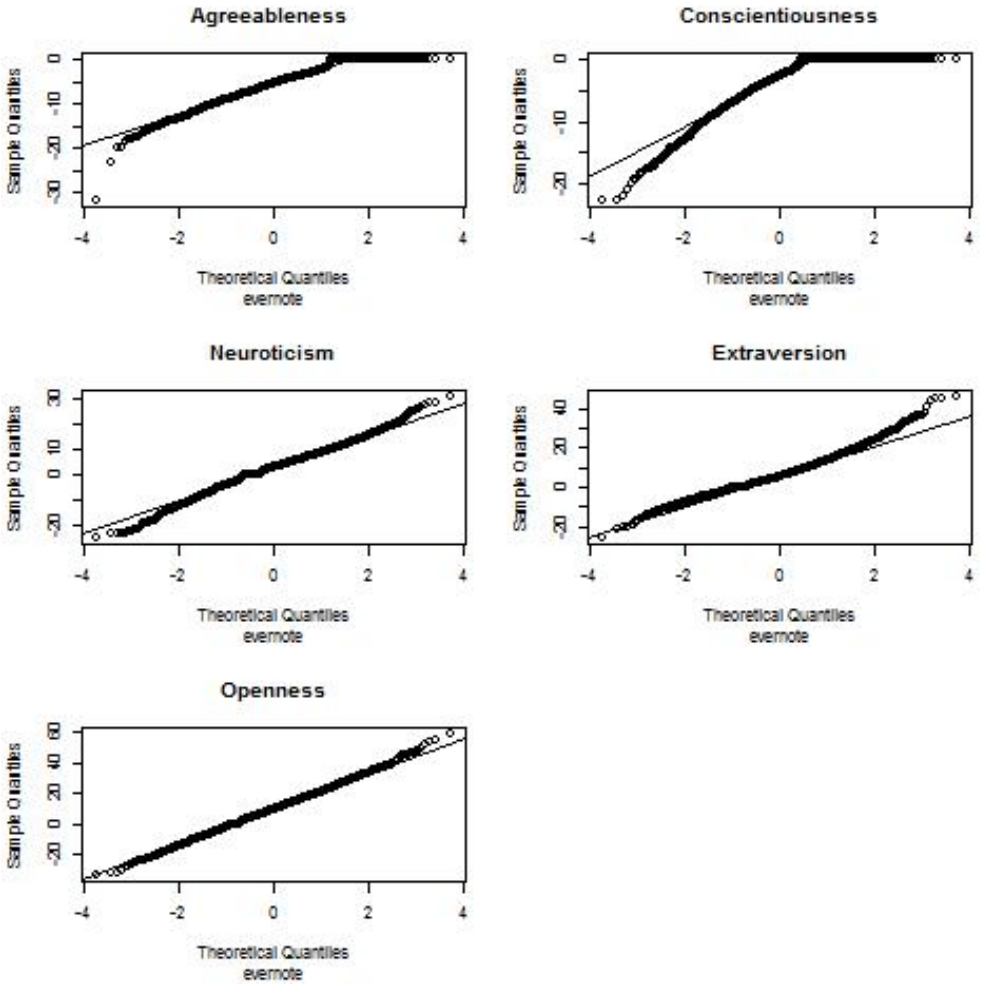


Figure 4.2.: QQ Plot der Daten von der App Evernote der Afrika-Länder

Anpassungstest, der die Verteilung der Stichprobe mit der Schiefe und Kurtosis der Normalverteilung abgleicht, um festzustellen, ob sich die Verteilungen gleichen. Die Hypothesen des Jarque-Bera-Tests lauten (vereinfacht):

H_0: Die Stichprobe ist normalverteilt
H_1: Die Stichprobe ist nicht normalverteilt

Ist der p-Wert kleiner als ein vorher festgelegtes Signifikanzniveau (hier 0.05) wird die Nullhypothese verworfen, was bedeutet, dass die Daten nicht normalverteilt sind. Im Gegenzug kann man bei nicht Ablehnung der Nullhypothese nicht sicher annehmen, dass eine Normalverteilung vorliegt. Dieser Test ist folglich eher dazu geeignet eine Normalverteilung auszuschließen.

Die Ergebnisse der Jarque-Bera-Tests (im Appendix zu sehen) bestätigten die Vermutung, dass die Daten für keine Forschungsfrage normalverteilt sind. Im Fall von großen Datensätzen kann jedoch bei keiner zu starken Abweichung von der Normalverteilung, diese Bedingung ignoriert werden. Jedoch sollten dann weitere Voraussetzungen, wie beispielsweise die Varianzhomogenität, zutreffen. Aus diesem Grund wurden die Datensätze im Anschluss auch auf Varianzhomogenität untersucht.

Varianzhomogenität

Eine weitere Voraussetzung an die Daten ist, dass die Varianzen der Merkmale (abhängige Variable) in den einzelnen Gruppen gleich groß sind. Diese Bedingung wird auch Varianzhomogenität genannt. Für diese Zwecke wird der Brown-Forsythe-Test [Sac99] verwendet, der besonders robust ist, falls beispielsweise keine Normalverteilung vorliegt und der Datensatz einige Ausreißer beinhaltet. Die Hypothesen des Brown-Forsythe-Tests lauten folgendermaßen:

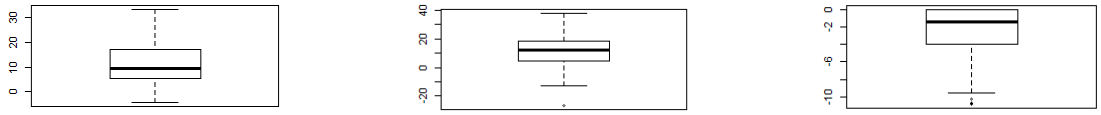
H_0: $\sigma_1^2 = \sigma_2^2 = .. = \sigma_k^2$
H_1: $\sigma_i^2 \neq \sigma_j^2$ mit i, j $\in \mathbb{N}$

Die Ergebnisse der Varianztests ergaben, dass für fast alle Forschungsfragen (alle außer Frage 2 bei Unterteilung der Daten nach Applikationen) die Merkmalsvarianzen der Stichproben unterschiedlich sind. Die Ergebnisse der Brown-Forsythe-Tests können im Appendix begutachtet werden [Sac99].

Gleiche Verteilungsform

Da der Datensatz der Studie weder normalverteilt ist, noch gleiche Populationsvarianzen vorweisen kann, sollten nichtparametrische Tests verwendet werden. Diese besitzen weniger Voraussetzungen als parametrische. Eine Bedingung, die bei einem in dieser Studie verwendeten Test existiert (Wilcoxon-Test), fordert dass die Zufallsstichproben aus Grundgesamtheiten ähnlicher Verteilungsformen stammen. Trifft diese Bedingung

Figure 4.3.: Boxplots von Verteilungen verschiedener Schiefen



Linksschief Symmetrisch Rechtsschief

zu, führt das zumindest zu einer Erhöhung der Genauigkeit des Tests [HS12].

Um die Gleichheit der Verteilungsformen der Stichproben zu überprüfen, können beispielsweise Boxplots verwendet werden. Boxplots sind die schnellste Methode einen Überblick über die Verteilung eines Datensatzes zu gewinnen. Vor allen die Schiefe der Daten ist leicht abzulesen. Bei symmetrischen Daten sollten die Whiskers der Boxplots ungefähr gleich lang sein und sich der Median (mittlerer Balken in der Box) in der Mitte der Box befinden. Bei rechts- oder linksschiefen Datensätzen ist ein Whisker kürzer als der andere, wobei sich der Median näher zu der Seite des kürzeren Whiskers befindet.

Letztendlich wurde mit Hilfe einer Vielzahl an Boxplots festgestellt, dass die Daten der Studie ausreichend ähnliche Verteilungsformen besitzen. Diese Boxplots befinden sich im Appendix. Abschließend ist festzuhalten, dass die Daten dieser Studie nicht ausreichend normalverteilt sind, ungleiche Varianzen besitzen und einer ähnlichen Verteilungsform folgen. Außerdem sind die Stichproben unabhängig von einander. Mit Hilfe all dieser Informationen können geeignete Tests für die statistischen Analysen ausgewählt werden.

4.7.2. Kruskal-Wallis-Test

Für die Forschungsfragen 1, 2, 3 und 5 werden Varianzanalysen benötigt. Es soll festgestellt werden, **ob** signifikante Unterschiede zwischen einzelnen Gruppen bestehen oder ob diese der selben Grundgesamtheit entstammen und **welche** Gruppen sich signifikant von einander unterscheiden. Zunächst muss die Existenz signifikanter Unterschiede überprüft werden. Die bekannteste Methode hierfür ist der Anova-Test. Da Anova ein parametrischer Test ist, wurde eine nichtparametrische Alternative ausgewählt: der Kruskal-Wallis-Test [HS12]. Mit Hilfe des Kruskal-Wallis-Tests wird der Einfluss einer unabhängigen Variable (z.B. kommentierte Applikation oder Herkunftsland des Reviewers) auf eine abhängige Variable (immer eine der Big Five Personality Traits) überprüft. Die unabhängige Variable bestimmt die verschiedenen "Gruppen", deren Unterschiede in der abhängigen Variable gemessen werden. Der Kruskal-Wallis-Test gibt an, ob unterschiedliche Messwerte der Gruppen signifikant sind. Signifikanz bedeutet in dem Zusammenhang, dass die Unterschiede nicht zufällig entstanden sind [HS12].

Verglichen mit den parametrischen Alternativen, besitzt der Kruskal-Wallis-Test eine asymptotische Effizienz von 95%. Er testet die Nullhypothese, dass alle k Gruppen aus

der gleichen Grundgesamtheit (Population) stammen. Die Alternativhypothese besagt, dass sich mindestens zwei der Gruppen signifikant von einander unterscheiden: [HS12].

H_0: Gruppen sind identisch
H_1: Gruppen sind nicht identisch

Als Signifikanzniveau wird bei allen Kruskal-Wallis-Tests ein Alpha von 0.05 gewählt. Alpha bezeichnet den Fehler 1. Art, also die Wahrscheinlichkeit die Nullhypothese fälschlicherweise abzulehnen. Die p-Werte (wahre Irrtumswahrscheinlichkeiten), die mittels des Kruskal-Wallis-Tests berechnet werden, müssen kleiner als das gewählte Alpha-Niveau sein (hier 0.05), damit die Nullhypothese verworfen werden kann und man von **signifikanten Unterschieden zwischen den Gruppen** sprechen kann [HS12].

Der Ausdruck "signifikanter Unterschied" ist nicht so zu verstehen, dass ein besonders großer Unterschied besteht. Die Prüfung auf Signifikanz bedeutet, dass untersucht wird, ob Unterschiede in den Stichprobenwerten (Gruppenwerten für abhängige Variable) zufällig entstanden sind oder ob es die gemessenen Unterschiede gibt, weil ein tatsächlicher Unterschied besteht. Wird die H_0-Hypothese abgelehnt, bedeutet das, dass es mindestens einen Unterschied zwischen den Stichprobenwerten gibt, der nicht zufällig entstanden ist. Dieser kann jedoch auch sehr gering sein.

4.7.3. Wilcoxon-Rangsummen-Test

Mit Hilfe des Kruskal-Wallis-Tests kann überprüft werden, ob signifikante Unterschiede zwischen bestimmten Stichproben bestehen. Beispielsweise für Forschungsfrage 1, ob in den Daten von Singapur die Werte für Neuroticism bei allen acht Applikationen aus der selben Population stammen. Wird ein signifikanter Unterschied identifiziert, ist es interessant zu erfahren, zwischen welchen Gruppen diese Unterschiede bestehen. Hierfür muss man einen sogenannten Post-Hoc-Test verwenden.

Bei Post-Hoc-Tests werden alle Gruppen paarweise, durch mehrmaliges Anwenden eines Tests für zwei unabhängige Stichproben, miteinander verglichen. Diese wiederholte Anwendung eines zwei-Stichproben-Tests führt jedoch zu einer sogenannten Alpha-Fehler-Kumulierung, also einer Erhöhung des Alpha-Fehlers. Um diesem Effekt entgegen zu wirken, haben die meisten Post-Hoc-Tests auch eine Alpha-Fehler-Korrektur integriert. Die zwei bekanntesten Methoden sind die Bonferronie- und die Holm-Korrektur [Sac99].

In dieser Studie wird ein paarweise (mehrfach) angewendeter Wilcoxon-Rangsummen-Test [HS12] mit Holm-Korrektur verwendet. Dieser Test ist auch unter dem Namen Mann-Whitney-U-Test bekannt. Der Wilcoxon-Test ist ein nicht parametrischer Test, der lediglich eine annähernd gleiche Verteilungsform der Stichproben voraussetzt (Siehe 4.7.1). Der bekannteste parametrische Test für diese Zwecke ist der Tukey-Test [HS12].

Die Nullhypothese des Wilcoxon-Tests besagt, dass die Stichproben gleiche Verteilungen besitzen. Die Alternativhypothese lautet hingegen, dass Unterschiede zwischen den Verteilungen existieren. Bei Ablehnung der Nullhypothese kann von signifikanten

Unterschieden ausgegangen werden. Die asymptotische Effizienz des Wilcoxon-Tests im Vergleich zu parametrischen Tests liegt bei 95%.

4.7.4. Spearmans Rangkorrelationskoeffizient

In den vorhergehenden Abschnitten wurden die Methoden vorgestellt, die angewendet werden, um die Forschungsfragen 1 bis 4 zu beantworten. Forschungsfrage 5 benötigt keine Varianzanalyse, sondern einen Korrelationstest, da die Korrelation zwischen zwei Merkmalen analysiert werden soll. Der Spearman Rangkorrelationskoeffizient r_s [Sac99] wurde als geeignetes Maß für die Korrelation ausgewählt. Im Anschluss an die Berechnung des Spearman Korrelationskoeffizienten wird auch die Signifikanz der Werte überprüft (Nullhypothese $r_s = 0$).

Der bekannteste Korrelationskoeffizient ist der nach Bavais und Pearson. Spearmans r_s ist im Vergleich zu dieser Methode die robustere Variante. Die Wirkung von Ausreißern wird abgeschwächt und die beiden Messreihen, deren Korrelation berechnet wird, müssen nicht normalverteilt sein. Außerdem wird bei Bavais-Pearson ein linearer Zusammenhang zwischen der beiden Merkmalen vorausgesetzt. [Sac99]

Für die Interpretation des Spearman Rangkorrelationskoeffizienten gilt es folgendes zu beachten: r_s befindet sich im Intervall [-1, 1], wobei -1 eine maximale negative Korrelation darstellt (je größer x, desto kleiner y oder je kleiner x desto größer y), 0 keine Korrelation bedeutet und 1 als maximale positive Korrelation zu interpretieren ist (je größer x, desto größer y oder je kleiner x desto kleiner y). Es gelten folgende Definitionen [Hau12]:

$r_s < 0$ bedeutet negativ korreliert
$r_s > 0$ bedeutet positiv korreliert
$r_s = 0$ bedeutet nicht korreliert
$0 \leq |r_s| < 0.5$ bedeutet schwach korreliert
$0.8 \leq |r_s| \leq 1$ bedeutet stark korreliert

Im folgenden Abschnitt werden die Ergebnisse der statistischen Analysen als Antworten auf die Forschungsfragen präsentiert.

5. Ergebnisse

In diesem Kapitel werden die Ergebnisse der Analysen für alle fünf Forschungsfragen zusammengefasst und Besonderheiten herausgearbeitet.

Die statistischen Analysen für die Forschungsfragen 1, 2 und 3 wurden für jede Persönlichkeitseigenschaft separat durchgeführt. Diese Tests wurden auf den gesamten Datensatz angewendet und je nach Forschungsfrage noch zusätzlich auf die Daten getrennt nach Ländern oder Applikationen.

Im Appendix befindet sich eine Vielzahl an Tabellen, die die Ergebnisse der statistischen Tests darstellen. Die Tabellen für die Forschungsfragen 1 bis 3 werden folgendermaßen interpretiert: Gehören alle Gruppen einer Analyse (die Gruppen entstehen durch die Unterteilung der Daten entsprechend der unabhängigen Variablen) auf Grund der Test-Ergebnisse des Kruskal-Wallis-Tests der gleichen Population an, wird diese Population nicht in der Tabelle aufgeführt. Beispielsweise hat ein Kruskal-Wallis-Test für die Frage 1 (mit Unterscheidung nach Ländern) ergeben, dass sich im Singapur-Datensatz die Messwerte für Openness der acht Apps (als die acht Gruppen) nicht signifikant unterscheiden und diese einer gemeinsamen Population angehören. Daher gibt es in Tabelle A.4 keinen Eintrag für Openness in Singapur. Bei den restlichen Big Five hatten sich immer mindestens zwei Gruppen signifikant unterschieden, weshalb diese auch in der Tabelle aufgeführt werden.

Nach signifikanten Kruskal-Wallis-Tests zeigen die Wilcoxon-Tests für jede Eigenschaft an, welche konkreten Gruppen der gleichen Population angehören und welche sich von einander unterscheiden. Diese Ergebnisse werden durch die Hochzahlen dargestellt. Die Hochzahlen müssen vertikal gelesen werden, da sie das Verhältnis der Gruppen für eine Persönlichkeitseigenschaft angeben. Besitzen zwei oder mehrere Gruppen die gleiche Hochzahl (bei mehreren Hochzahlen genügt eine gemeinsame) bedeutet das, dass diese bezüglich der Eigenschaft keine signifikanten Unterschiede von einander aufweisen und der gleichen Population angehören. Eine Gruppe kann auch in mehreren Populationen enthalten sein. Signifikant unterschiedliche Gruppen sind daran zu erkennen, dass sie keine gemeinsame Hochzahl besitzen. Die Hochzahlen geben außerdem die Rangfolge der Populationen an. Die Population mit der kleinsten Hochzahl besitzt den geringsten und die Population mit der größten Hochzahl den höchsten Median der entsprechenden Persönlichkeitseigenschaft. Die Median-Werte dienen ausschließlich dem Vergleich der einzelnen Populationen und sind nicht absolut zu sehen.

5.1. Einfluss der Applikation auf die Eigenschaftswerte

Forschungsfrage 1 lautet **„Unterscheiden sich die Persönlichkeiten der Reviewer abhängig von der Applikation?“**. Zur Beantwortung dieser Frage wurde der Einfluss der unabhängigen Variablen „Applikation“ auf jeweils eine der Big Five Personality Traits als abhängige Variable untersucht. Diese Analysen wurden einerseits über dem gesamten Datensatz (Siehe Tabelle A.1) und andererseits über den nach Ländern getrennten Datensätzen (Siehe Tabellen A.2 bis A.6) ausgeführt.

Die Übersetzungen zu den englischen Begriffen der Big Five sind im Glossar nachzulesen. Die Aussagen in dieser Studie über die User der acht ausgewählten Applikationen und der verschiedenen Länder sind nur Vermutungen die auf Grund der Analyse von bestimmten Userkommentaren eines bestimmten Zeitraums entstanden sind. Sie haben keine Aussagekraft über die Allgemeinheit der User der Applikationen aus den verschiedenen Ländern.

Als erstes wurden die Ergebnisse der Kruskal-Wallis-Tests überprüft. Bei der Anwendung der Tests auf den gesamten Datensatz waren alle Ergebnisse signifikant und bei der Anwendung auf die nach Ländern getrennten Datensätze 24 von 25 Stück. Das bedeutet, dass bei fast allen Analysen mindestens ein signifikanter Unterschied zwischen den Gruppen identifiziert wurde. Das genaue Verhältnis der Gruppen ist in den Tabellen A.1 bis A.6 zu sehen. Einige Besonderheiten der Ergebnisse, die in diesen Tabellen visualisiert werden, werden nun herausgearbeitet.

Zunächst wurde ermittelt bei welchen Persönlichkeitseigenschaften besonders viele signifikante Unterschiede bestehen. Dies wurde auf Grund der Anzahl an Populationen für jede Eigenschaft bestimmt (mehr Populationen, also mehr verschiedene Hochzahlen, bedeutet mehr Unterschiede). Bei der Analyse des gesamten Datensatzes (Tabelle A.1) konnten die meisten Unterschiede bei den Eigenschaften Openness, Extraversion und Conscientiousness identifiziert werden. Die Analysen, der nach Ländern aufgeteilten Daten, bestätigten diese Ergebnisse. Die Tabellen A.2 bis A.6 zeigen an, dass bei den Daten von den Vereinigten Staaten und den Afrika-Ländern durchschnittlich die meisten Unterschiede der Gruppen (gemessen in der durchschnittlicher Populationsanzahl) existieren. Singapur hingegen weist die geringsten Unterschiede zwischen den Gruppen auf.

Daraufhin wurde überprüft, welche Applikationen im Vergleich zu den anderen die Kommentare mit den höchsten oder niedrigsten Medianwerten einer Eigenschaft besitzen. Bei den Analysen des gesamten Datensatzes (Tabelle A.1) konnte folgendes festgestellt werden: Am wenigsten neurotisch, also die höchste emotionale Stabilität, sowie der höchste Wert an Gewissenhaftigkeit, konnte bei den Kommentaren der User der Booking App identifiziert werden. Am neurotischsten und unverträglichsten (niedrige Agreeableness) ist das Feedback der Angry Birds User. Deren Reviews besitzen außerdem den niedrigsten Wert an Gewissenhaftigkeit. Das Feedback der Tripadvisor User deutet daraufhin, dass diese besonders offen sind. Am wenigsten Offenheit wurde bei Pinterest- und WhatsApp-Kommentaren entdeckt, was bedeutet, dass diese Reviewer eher konventionell sein könnten. Die WhatsApp Reviews besitzen außerdem die höch-

sten Werte an Verträglichkeit/ Freundlichkeit. Die Kommentare von Pinterest sind am extrovertiertesten und die von Adobe Reader am introvertiertesten. Die Applikationen Dropbox und Evernote haben im Vergleich zu den anderen Applikationen immer eher mittlere Eigenschaftswerte.

Die Extremas der Medianwerte (höchste und niedrigste Werte) der einzelnen Persönlichkeitseigenschaften bei der Aufteilung nach Ländern unterscheidet sich nur geringfügig von den Ergebnissen in der Tabelle A.1. Außerdem sind sich die Extremas zwischen den verschiedenen Ländern auffällig ähnlich.

Auffallend bei allen Analysen ist, dass oft vor allem folgende Applikationen häufig zu einer gemeinsamen Population gehören: Adobe Reader, Dropbox und Evernote (vor allem bei Neuroticism, Agreeableness und Openness). Diese Applikationen gehören den gleichen oder zumindest ähnlichen App-Kategorien an. Dropbox und Evernote sind der Kategorie „Productivity“ zuzuordnen, die sich thematisch nicht stark von der Kategorie „Business“ von Adobe Reader unterscheidet. Vor allem die Booking App und Adobe Reader zeigen wenige Unterschiede zu anderen Gruppen auf.

Insgesamt wurden viele signifikante Unterschiede (viele verschiedene Populationen und 24 von 25 signifikante Kruskal-Wallis-Tests) abhängig von der Applikation identifiziert. Nach Sichtung der Ergebnisse liegt jedoch auch die Vermutung nahe, dass sich die applikationsbedingten Unterschiede in den Eigenschaftswerten nicht besonders stark nach Herkunftsland der Reviewer unterscheiden, da die Populationsränge der Tabellen A.2 bis A.6 sehr ähnlich verteilt sind.

5.2. Abhängigkeit der Big Five von dem Herkunftsland der Reviewer

Nun sollen die Ergebnisse der zweiten Forschungsfrage vorgestellt werden. Diese lautet **Unterscheiden sich die Persönlichkeiten der Reviewer abhängig von deren Herkunft?**. Bei den statistischen Tests wurde die Abhängigkeit der Big Five von der unabhängigen Variablen "App Store" analysiert. In Tabelle A.7 sind die Ergebnisse der Analysen des gesamten Datensatzes zu sehen. Die Tests wurden ebenfalls für die Daten jeder Applikation separat durchgeführt (Siehe Tabellen A.8 bis A.13).

Als erstes wurden die Kruskal-Wallis-Tests für den gesamten Datensatz ausgewertet. Diese fünf Stück waren signifikant. Bei der Auswertung der Tests, die für die Daten jeder Applikation separat durchgeführt wurden, waren hingegen nur noch 19 von 40 Kruskal-Wallis-Tests signifikant. Bei den Daten der Applikationen Adobe Reader und Booking konnten gar keine landesspezifischen Persönlichkeitseigenschaften identifiziert werden und bei den restlichen Applikationsanalysen waren vor allem die Eigenschaften Extraversion und Agreeableness selten vom Herkunftsland des Reviewers abhängig. Die signifikanten Unterschiede, die identifiziert werden konnten, sind in den Tabellen A.7 bis A.13 zu sehen. Bei den Applikationen Evernote, Tripadvisor, Pinterest und WhatsApp wurden die meisten landesspezifischen Unterschiede der Big Five Werte entdeckt.

Zu Tabelle A.7 bedarf es einer genaueren Erläuterung. Die Eigenschaft Neuroticism wird hier nicht aufgeführt, nicht weil in der Eigenschaft keine signifikanten Unterschiede bestehen, sondern weil es nicht möglich war die Ergebnisse auf gewohnte Weise darzustellen. Der Wilcoxon-Test gibt an, dass sich alle Gruppen im Neuroticism-Wert signifikant von einander unterscheiden, außer die Afrika-Länder und die Vereinigten Staaten, die der selben Population angehören. Jedoch sind die Median-Werte aller Länder außer Großbritannien gleich groß, weshalb es nicht möglich ist eine Rangfolge der Hochzahlen zu bilden. Das bedeutet nicht, dass die Ergebnisse des Wilcoxon-Tests falsch sind. Denn dieser vergleicht Rangsummenvarianzen und nicht die Mediane. In den Tabellen werden die Mediane dargestellt, weil diese am besten die Unterschiede, die der Wilcoxon-Test misst, visualisieren können. Nur in diesem Fall war die Visualisierung (von den Unterschieden der verschiedenen Regionen im Wert von Neuroticism) nicht möglich.

Bei dieser Forschungsfrage konnten nur wenige Unterschiede der Eigenschaftswerte zwischen den verschiedenen Gruppen ermittelt werden. Meistens gibt es pro Eigenschaft zwei Populationen, die unterschiedlich zu einander sind. Am meisten landesspezifische Unterschiede zwischen den Gruppen wurden bezüglich der Persönlichkeitseigenschaften Neuroticism und Openness in den Datensätzen von Pinterest und WhatsApp ermittelt.

In Tabelle A.7 ist bereits zu erkennen, dass sich die Gruppen Afrika-Länder und Vereinigte Staaten nie signifikant unterscheiden und daher immer zu der selben Population gehören. Bei den Eigenschaften Extraversion und Conscientiousness existieren keine Unterschiede zwischen Großbritannien und Kanada. Bei der Unterteilung nach Applikationen (Siehe Tabelle A.7 bis A.13) werden folgende Zusammenhänge deutlich: Auch hier unterscheiden sich die Big Five von den Afrika-Ländern und den Vereinigten Staaten nie. Bezüglich der Eigenschaften Openness und Neuroticism gehören Großbritannien, Singapur und Kanada oft zu der gleichen Population. Der einzige auffällige Unterschied besteht zwischen der Population Afrika-Länder/Vereinigte Staaten und der Gruppe Großbritannien. Die Eigenschaftswerte der Kommentare aus Singapur weisen nur sehr selten Unterschiede zu den Big Five Werten der anderen Regionen auf.

Als nächstes wurde untersucht, welche Regionen die Kommentare mit den höchsten oder niedrigsten Medianwerten der einzelnen Persönlichkeitseigenschaften aufweisen. Hierfür wurden zunächst die Ergebnisse der Analyse des gesamten Datensatzes überprüft (Siehe Tabelle A.7), wobei folgendes festgestellt wurde: Das Feedback der amerikanischen und afrikanischen (aus den Ländern der Afrika-Gruppe) Reviewer besitzt die geringsten Werte an Offenheit und die höchsten Werte an Extraversion und Conscientiousness. Am wenigsten Conscientiousness und Extraversion konnte bei den Kommentare aus Singapur festgestellt werden. Die Reviews aus Kanada waren im Vergleich zu den anderen am wenigsten freundlich und die Kommentare aus Großbritannien befanden sich für alle fünf Persönlichkeitseigenschaften im mittleren Bereich. Leider konnte der Vergleich von Neuroticism nicht aufgestellt werden. Eine ähnliche Untersuchung der Mediane der Tabellen A.7 bis A.13) wäre nicht besonders aufschlussreich, da die landesspezifischen Unterschiede der Applikationen nur sehr gering sind.

Zusammenfassend lässt sich festhalten, dass landesspezifische Unterschiede in den Big Five Werten bestehen (signifikante Kruskal-Wallis-Tests beim gesamten Datensatz). Jedoch gibt es nur wenige landesspezifischen Unterschiede innerhalb der Applikationen. Bei den Applikationen, wo Abhängigkeiten vom Herkunftsland der Reviewer identifiziert wurden, waren am auffälligsten die Gemeinsamkeiten zwischen den Afrika-Ländern und Vereinigte Staaten Reviews und deren Unterschied zu dem Feedback aus Großbritannien.

5.3. Zusammenhang der Big Five Werte mit der Sterne Bewertung der User

Als nächstes wurde die dritte Forschungsfrage analysiert. Die Forschungsfrage 3 „Unterscheiden sich die Persönlichkeiten der Reviewer, die eine positive, neutrale oder negative Sterne-Bewertung abgegeben haben?“ untersucht den Einfluss der Bewertung des Kommentars durch die Sterne Vergabe auf die Persönlichkeitseigenschaften der Reviewer. 1-2 Sterne wurden als negativer Kommentar, 3 Sterne als neutraler und 4-5 Sterne als positiver Kommentar substituiert. Es soll herausgefunden werden, ob zwischen den beiden Aspekten ein Zusammenhang besteht. Zunächst wurden die statistischen Tests mit den gesamten Daten der Studie durchgeführt (alle Länder und alle Apps). Die Ergebnisse können in Tabelle A.14 betrachtet werden. Die bewertungsbedingten Unterschiede der Big Five wurden auch für die einzelnen Applikationen getrennt berechnet (Tabelle A.15 bis A.22).

Die Kruskal-Wallis-Tests, die die Abhängigkeit der Persönlichkeitseigenschaften von den drei Kategorien "negative Sterne-Bewertung", "neutrale Sterne-Bewertung" und "positive Sterne-Bewertung" des gesamten Datensatzes ermitteln sollten, waren alle signifikant. Bei der Analysen, der nach verschiedenen Applikationen unterteilten Daten, wurde festgestellt, dass 36 von 40 Tests signifikant waren. Nur die Analysen der Persönlichkeitseigenschaften Openness und Agreeableness der Applikationen Adobe Reader und Booking offenbarten keine signifikant unterschiedlichen Gruppen.

Die Anzahl an Populationen innerhalb einer Eigenschaft ist höher, je mehr Unterschiede zwischen den einzelnen Gruppen bestehen. Bei dieser Forschungsfrage ist maximal eine Anzahl von 3 Populationen möglich, falls sich alle drei Gruppen von einander unterscheiden. In Tabelle A.14 kann gesehen werden, dass die Analysen meist drei verschiedene Populationen ergeben haben. Bei den Analysen der Eigenschaften getrennt nach Applikationen gibt es am meisten Unterschiede in den Big Five Werten der Persönlichkeitseigenschaften Neuroticism, Extraversion und Conscientiousness.

Als nächstes wurden die Extremas der Mediane identifiziert. Die Ergebnisse der Analyse des gesamten Datensatzes bedeuten folgendes: Die Kommentare, denen vier oder fünf Sterne vergeben wurden, die folglich positiv sind, besitzen die niedrigsten Werte der Persönlichkeitseigenschaften Neuroticism und Openness. Den Analysen zu Folge sind User dieser Kommentare jedoch am freundlichsten, extrovertiertesten und gewissenhaftesten. Das negative Feedback wurde als besonders neurotisch und offen

identifiziert. Außerdem wurden die niedrigsten Werte für die Eigenschaften Extraversion und Conscientiousness entdeckt. Die neutralen Reviews lassen auf besonders offene und am wenigsten freundliche User schließen. Die Analysen der nach Applikationen getrennten Datensätze führten vor allem bei den Eigenschaften Neuroticism, Extraversion und Conscientiousness zu besonders einheitlichen Ergebnisse (bei fast allen Apps gleich), die denen in Tabelle A.14 gleichen. Die Big Five Werte der restlichen Persönlichkeitseigenschaften waren stärker von den verschiedenen Applikationen abhängig.

Die meisten Gemeinsamkeiten wurden zwischen den neutralen und negativen Kommentaren entdeckt. Vor allem das positive und neutrale Feedback offenbarte die häufigsten Unterschiede.

Alles in allem wurden viele Unterschiede in den Big Five Werten abhängig von der Sterne-Bewertung identifiziert (viele signifikante Kruskal-Wallis-Tests und viele Populationen). Teilweise unterschieden sich diese auch zwischen den einzelnen Applikationen. Hierbei gab es die wenigsten Gemeinsamkeiten zwischen dem neutralen und negativen Feedback.

5.4. Korrelation der Sentiments mit den vergebenen Sternen

Die Ergebnisse der beiden folgenden Forschungsfragen beziehen sich auf die Sentiments der User. Die vierte Forschungsfrage lautet "Korrelieren die abgegebenen Sterne-Bewertungen mit den Sentiments der Reviewer?". Eine positive Korrelation würde bedeuten, dass eine größere Anzahl an Sternen auch einer positiveren Stimmung des User entspricht und umgekehrt eine niedrigere Sterne-Bewertung auf eine negativere Stimmung schließen lässt. Dadurch soll festgestellt werden, ob die Stimmungen, genau wie die Sterne-Bewertungen, als die Zufriedenheit der User mit den Applikationen interpretiert werden können.

Zunächst wurde der gesamte Datensatz auf Korrelationen mit den Sterne-Bewertungen untersucht. Da die identifizierten Korrelationen teilweise recht niedrig waren, wurden bei weiteren Analysen auch die Korrelationen der negativen Sentiments ($(-1)*$ "Negative Emotion" zugewiesen) und positiven Sentiments ("Positive Emotion" zugewiesen) mit der Sterne Vergabe ermittelt, um mögliche Unterschiede zu überprüfen. All diese Untersuchung wurden außerdem noch für die Daten jeder Applikation separat durchgeführt.

Bei der Analyse des gesamten Datensatzes wurde ein signifikanter Spearman- Rangkorrelationskoeffizient von 0.4784 berechnet. Dieser Wert zeigt eine positive Korrelation zwischen den Sterne-Bewertungen und den Sentiments der Kommentare auf. Der Wert stellt, auf eine Nachkommastelle gerundet, per Definition (Kapitel 4.7.4) eine mittel starke Korrelation dar. Bei den nach positiven und negativen Sentiments getrennten Analysen, wurden signifikante, schwache Korrelationen von 0.3657 und 0.0507 identifiziert. Vor allem die Korrelation der negativen Emotionen ist nur sehr niedrig, jedoch bestätigen die Ergebnisse den vermutete Zusammenhang: Je höher die positiven oder negativen Sentiments, desto größer die Sterne-Anzahl. Diese Ergebnisse werden in den

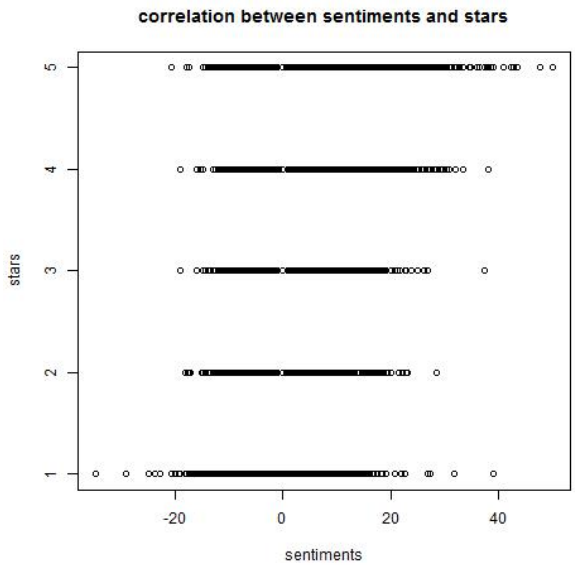


Figure 5.1.: Korrelation der Sentiments mit den Sterne-Bewertungen

Abbildungen 5.1, 5.2 und 5.3 visualisiert.

Die Korrelationen der Sterne-Bewertungen mit allen Sentiments, den positiven und negativen Sentiments für die verschiedenen Applikationen sind in Tabelle A.23 bis A.24 zu sehen. Alle signifikanten Werte sind fett gedruckt.

Bei der Analyse der Sentiments der einzelnen Applikationen wurden ebenfalls schwach positive Korrelationen ermittelt, welche sich in dem Intervall $[0.38; 0.49]$ befinden. Alle Werte sind signifikant, wobei die Booking App den niedrigsten und Tripadvisor und WhatsApp die höchsten Rangkorrelationskoeffizienten aufweisen. Die Korrelationen der positiven Sentiments sind ebenfalls alle signifikant und schwach positiv. Bei den Analysen der Korrelationen der negativen Sentiments mit den Sterne-Bewertungen wurden nur drei signifikante Werte identifiziert. Diese wurden bei den Applikationen Evernote, Tripadvisor und WhatsApp entdeckt. Diese Rangkorrelationskoeffizienten sind sehr schwach positiv und befinden sich in dem Intervall $[0.10; 0.19]$. Die Plots zu den einzelnen Korrelationen sind im Appendix zu sehen.

Letztendlich wurden sowohl beim gesamten Datensatz, als auch bei den einzelnen Applikationen viele signifikante, schwach positive Korrelationen ermittelt. Jedoch scheinen vor allem die negativen Korrelationen nur sehr niedrig zu sein und sind bei manchen Applikationen auch nicht signifikant.

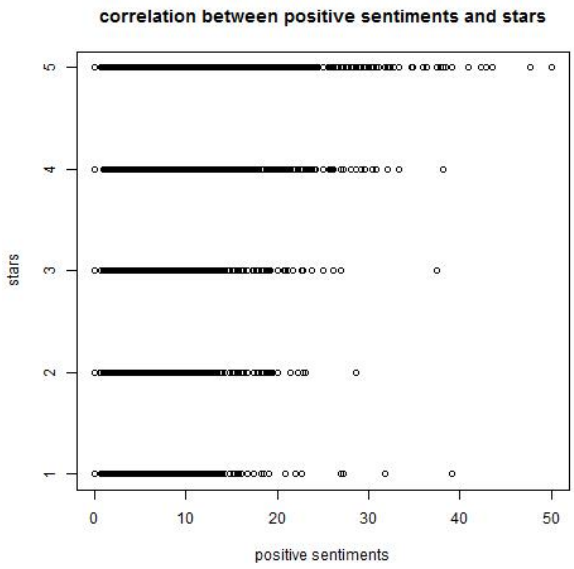


Figure 5.2.: Korrelation der positiven Sentiments

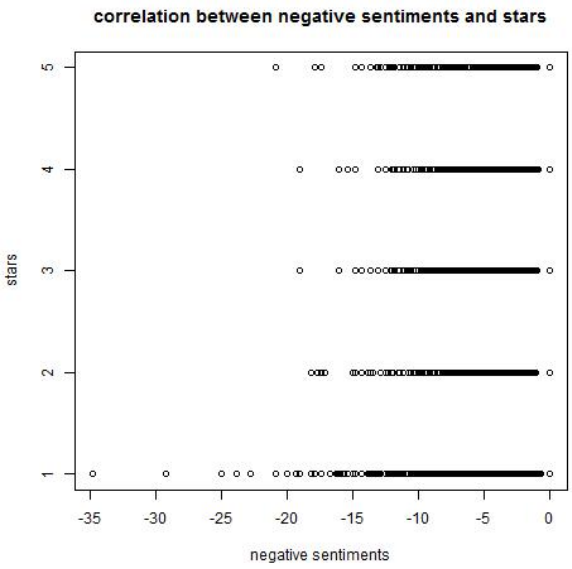


Figure 5.3.: Korrelation der negativen Sentiments

5.5. Veränderungen der User-Sentiments bei verschiedenen Versionen eine App

Bei der Analyse der Forschungsfrage "Unterscheiden sich die Sentiments der Reviewer innerhalb einer Applikation abhängig von der Versionsnummer?" wurde der Einfluss der unabhängigen Variable Versionsnummer auf die Sentiments in den Reviews untersucht. Die Ergebnisse der Wilcoxon-Tests für die Applikationen Angry Birds und Booking sind in den Tabellen A.25, A.26, A.28 und A.29 im Appendix zu sehen. Die Mediane der Sentiments für die einzelnen Versionsnummer befinden sich in den Tabellen A.27 und A.30.

Acht von acht Kruskal-Wallis-Tests (einen für jede Applikation) waren signifikant, das heißt, dass bei jeder App mindestens ein signifikanter Unterschied der Sentiments zwischen zwei Versionsnummern gemessen wurde. Die Werte, die auf signifikante Unterschiede schließen lassen wurden fett gedruckt. Das bedeutet, dass es tatsächlich möglich ist Stimmungsänderungen abhängig vom Release zu messen. In der Tabelle A.25 bis A.26 unterscheiden sich beispielsweise die Versionen 5.3 und 5.8 signifikant von einander. Bei der Version 5.3 wurde der Median-Wert 11.370 und bei Version 5.8 der Wert 5.810 bestimmt. Das heißt, dass die Stimmung der User bei der neueren Version schlechter ist als davor.

In Tabelle A.28 bis A.29, die die Unterschiede in den Stimmungen der Angry Birds-User abhängig von der Versionsnummer aufzeigt, sind mehr signifikante Unterschiede zu erkennen. Von Version 2.2.0 auf 2.3.0 ist ebenfalls eine Verschlechterung der Stim-

mung zu sehen. Interessant ist auch die Veränderung der Sentiments von Version 3.1.0 auf 3.1.2, da diese Versionen direkt auf einander folgen. Bei 3.1.2 kam es zu einer Verbesserung der Stimmung der User, was bedeutet, dass die User mit dem neuen Release zufriedener waren.

Ist der Unterschied der Sentiments zwischen zwei Versionen nicht signifikant, bedeutet das nicht, dass definitiv kein Veränderung der Stimmungen besteht, sondern dass die Wahrscheinlichkeit, dass der gemessene Unterschied nicht zufällig entstanden ist, zu gering ist. Es können in so einem Fall also keine eindeutigen Schlüsse gezogen werden, es ist jedoch zu vermuten, dass es möglicherweise keine Änderung der Stimmungen gegeben hat.

Auf diese Weise kann der Hersteller einer App überprüfen, ob und wie sich die Stimmungen der User von Version zu Version verändert haben. Eine weitere Möglichkeit die Sentiment Analyse einzusetzen, wäre beispielsweise die Stimmungen der User bezüglich verschiedenen Features zu bestimmen. [**Guzman**]

6. Interpretation und Future Work

In dieser Section werden die Ergebnisse der Forschungsfragen interpretiert. Das bedeutet für einige Zusammenhänge werden Vermutungen für mögliche Ursachen angestellt.

6.1. Applikationsabhängige Unterschiede

Als erstes werden die Ergebnisse der ersten Forschungsfrage interpretiert. Den Ergebnissen der Forschungsfrage 1 im Abschnitt 5.1 zu Folge besteht eine Abhängigkeit zwischen den Applikationen und den Persönlichkeitseigenschaften der User. Die Eigenschaften der User unterscheiden sich folglich signifikant von einander, je nachdem welche App diese verwenden. Das kann bedeuten, dass sich verschiedene Typen von Menschen zu verschiedenen Applikationen hingezogen fühlen. In diesem Fall wäre es für den Erfolg einer Applikation notwendig, die applikationsspezifischen Persönlichkeitseigenschaften der User im Softwareentwicklungsprozess zu erforschen.

Im Zuge der Analyse der Ergebnisse wurde festgestellt, dass sich vor allem die User der Applikationen Evernote, Dropbox und Adobe Reader selten von einander unterscheiden. Das könnte ein Hinweis darauf sein, dass die User von Apps gleicher oder ähnlicher Kategorien weniger unterschiedlich sind. Evernote und Dropbox gehören beide zu der Kategorie Produktivität, Adobe Reader zu dem Genre Wirtschaft (Business). Diese beiden App Store Kategorien unterscheiden sich thematisch nicht stark oder sind sich zumindest von den Kategorien der ausgewählten Apps am ähnlichsten. Der Inhalt der beiden Genres ist nicht eindeutig definiert, jedoch scheint bei beiden Kategorien das Ziel zu sein, den Alltag der User zu erleichtern und dessen Organisation zu unterstützen. Für eine Analyse der Abhängigkeit der Persönlichkeitseigenschaften von dem Genre der Applikationen könnte in zukünftigen Arbeiten die Vorgehensweise bei den ersten drei Forschungsfragen als Roadmap dienen. Es würde die Analyseergebnisse deutlich verbessern, wenn für jede Kategorie die Daten mehrerer Applikationen verwendet werden würden (auch mehrere verschiedene Kategorien). Ein anderer Ansatz wäre als abhängige Variable den Hauptzweck der App zu verwenden. Also beispielsweise "Unterhaltung" für Angry Birds, "Organisiation und Verwaltung" für Evernote, Adobe Reader und Dropbox usw. Denn manchmal unterscheiden sich die Applikationen, die der selben Kategorie angehören in ihrem Zweck. Pinterest und WhatsApp gehören zum Beispiel zur Kategorie "Soziale Netzwerke". WhatsApp wird jedoch hauptsächlich zur Kommunikation benutzt und Pinterest mehr zur Selbstdarstellung und dem Austausch von Ideen und Interessen mit anderen Nutzern. Folglich werden die beiden Applikationen aus verschiedenen Intentionen genutzt. Und tatsächlich wurden keine besonderen

Zusammenhänge zwischen den Eigenschaften der User der zwei Apps entdeckt.

Die maximalen und minimalen Mediane der verschiedenen Applikationen bieten bereits eine Möglichkeit, Informationen über die Personas der Applikationen zu gewinnen. Die Ergebnisse aus Tabelle A.1 könnten beispielsweise folgendermaßen interpretiert werden: Die User von Pinterest und WhatsApp legen Wert auf die Pflegung von sozialen Kontakten. Die WhatsApp User sind besonders freundlich und die von Pinterest sehr extrovertiert, was beides für ein ausgeprägtes soziales Leben spricht. Außerdem scheinen die User beider Applikationen eher konventionell zu sein. Die Analysen haben ergeben, dass Angry Birds User geringe Werte an Conscientiousness und Agreeableness besitzen. Außerdem erzielten sie die höchsten Werte für Neuroticism. Man könnte vermuten, dass diese User gelangweilter und schlechter gelaunt sind, als die Nutzer der anderen Apps und deswegen mehr Zeit mit dem Spielen auf dem Smartphone verbringen. Die Analysen der User der Booking App lassen darauf schließen, dass diese sehr gewissenhaft und selbstbewusst veranlagt sind. Hierbei könnte man sich zum Beispiel einen erfolgreichen Geschäftsmann- oder Frau vorstellen, die geschäftlich viel reisen und daher die Booking App verwenden. Beim Feedback von Tripadvisor wurden hingegen vor allem hohe Werte von Offenheit identifiziert. Das lässt eher auf User schließen, die gerne reisen, um neue Kulturen kennen zu lernen und neue Erfahrungen zu machen und diese App während oder vor ihren Reisen verwenden.

Bei den Eigenschaftswerten von Extraversion, Conscientiousness und Agreeableness wurden die meisten applikationsabhängigen Unterschiede identifiziert. Das kann bedeuten, dass vor allem diese Persönlichkeitseigenschaften beeinflussen welche App ein User gerne verwendet. Des Weiteren wurden bei den Reviews der Afrika-Länder und der Vereinigten Staaten am meisten Unterschiede zwischen den Apps identifiziert. Bei Singapur hingegen konnten am wenigsten Unterschiede festgestellt werden. Eine mögliche Theorie, die bei zukünftigen Arbeiten überprüft werden könnte, ist dass die Anzahl der applikationsbedingten Unterschiede abhängig von der flächenmäßigen Größe des Landes des Reviewers ist. Es wäre nämlich denkbar, dass bei Singapur so wenige Unterschiede zwischen den Applikationen bestehen, da die User alle in der selben Stadt leben und sich deswegen eventuell ähnlicher sein könnten (z.B. auf Grund des Klimas). Die Vereinigten Staaten hingegen sind ein flächenmäßig großes Land und die Gruppe der Afrika-Länder besteht aus mehreren verschiedenen Ländern in Westafrika. Daher gibt es in diesen Regionen größere klimatische Unterschiede zwischen den Usern.

Letztendlich konnte die Hypothese, dass applikationsabhängige Unterschiede in den Eigenschaftswerten der User bestehen, bestätigt werden. Außerdem wurde ein Ansatz gezeigt, um bereits erste Informationen über die Personas der User einer App zu gewinnen.

6.2. Abhängigkeit der Big 5 von der Herkunft der Reviewer

Mögliche Ursachen für die Ergebnisse der statistischen Analysen, die in Abschnitt 5.1 vorgestellt wurden, werden nun dargestellt.

Al nächstes konnte festgestellt werden, dass landesspezifische Unterschiede in den Persönlichkeitseigenschaften der Reviewer in dem User Feedback dieser Studie existieren. Das bedeutet bei dem, in dieser Studie verwendeten User Feedback, wurde ein Einfluss des Herkunftslandes der Reviewer auf deren Eigenschaften identifiziert. Des Weiteren wurde festgestellt, dass sich innerhalb einer Applikation die Herkunft des Reviewers nur gering und nur bei manchen Applikationen auf dessen Eigenschaften auswirkt. Kombiniert man diese Erkenntnisse mit den Ergebnissen der bisherigen Forschungsfragen, könnte man folgendermaßen argumentieren: In den Reviews existieren zwar Unterschiede der User je nach Herkunftsland, jedoch benutzen trotzdem User mit ähnlichen Persönlichkeiten bevorzugt bestimmte Applikationen. Das wäre ein möglicher Grund weshalb die landesspezifischen Unterschiede innerhalb einer Applikation nur sehr gering ausfallen. Die Hypothese, dass sich die User aus verschiedenen Regionen bezüglich ihrer Persönlichkeitseigenschaften unterscheiden, konnte für den Datensatz dieser Studie bestätigt werden, auch wenn innerhalb der Applikationen nur wenige Unterschiede festgestellt wurden. Für die Softwareentwicklung von Apps könnten diese Ergebnisse konkret bedeuten, dass bei der Erschließung neuer internationaler Märkte wahrscheinlich keine Anpassung der Applikationen an die User dieser Länder durchgeführt werden muss. Jedoch sollte für jede Applikation extra überprüft werden, ob landesspezifische Unterschiede der User bestehen. In zukünftigen Arbeiten müsste auch untersucht werden, ob die Ergebnisse dieser Studie verallgemeinert werden können.

Bei der Analyse des gesamten Datensatzes wurde festgestellt, dass sich die Big Five Werte der Reviews den Afrika-Ländern und den Vereinigten Staaten immer gemeinsam in einer Population befinden. Auch Kanada und Großbritannien unterscheiden sich bezüglich der Eigenschaften Extraversion und Conscientiousness nicht von einander. Die Ergebnisse der Analyse der nach Applikationen getrennten Daten ergaben, dass sich bei den Apps, bei denen landesspezifische Unterschiede festgestellt wurden, meistens die Eigenschaften der Reviewer aus Großbritannien von denen der Population Afrika-Länder/Vereinigte Staaten unterscheiden. Zukünftige Studien könnten genauer erforschen, was die zu Grunde liegenden Ursachen für diese Entdeckungen sein könnten. Es könnte beispielsweise untersucht werden, ob die entsprechenden Kontinente der Länder Einfluss auf die Big Five Werte der Reviews haben. Oder ob sich insbesondere die Reviews der Länder unterscheiden, die in unterschiedlichen Breitengraden liegen. Der Unterschied der Großbritannien User zu denen aus den Afrika-Ländern und den Vereinigten Staaten würde diesen Theorien zumindest nicht widersprechen.

Die Eigenschaften Extraversion und Agreeableness zeigen besonders wenige landesspezifische Unterschiede innerhalb der Applikationen auf. Das könnte heißen, wenn eine Anpassung von Apps an neue Märkte durchgeführt werden muss, dann weil sich die User bezüglich der Eigenschaften Openness, Conscientiousness und Neuroticism unterscheiden.

Die Ergebnisse in Tabelle A.7 können bereits genutzt werden, um Unterschiede zwischen den Reviewern verschiedener Ländern zu erkennen. Beispielweise könnten die Reviewer aus den Vereinigten Staaten und den Afrika-Ländern dieser Studie im Vergleich zu den anderen Reviewern als besonders konventionell, extrovertiert und gewissenhaft bezeichnet werden. Noch interessanter sind jedoch die Tabellen A.8 bis A.13, die aufzeigen, welche landesspezifischen Unterschiede innerhalb einer App bestehen. Ein solcher Vergleich sollte bei der Internationalisierung einer Applikation für alle relevanten Regionen durchgeführt werden.

6.3. Einfluss der Sterne-Bewertungen auf Persönlichkeitseigenschaften

Bei der Analyse von Forschungsfrage 3 wurde untersucht, ob ein Zusammenhang zwischen den Persönlichkeitseigenschaften der Reviewer und deren Sterne-Bewertung existiert. Die signifikanten Kruskal-Wallis-Tests und die vielen verschiedenen Populationen (fast immer Unterschiede zwischen allen drei Gruppen), die die Analyse des gesamten Datensatzes und der Daten der einzelnen Applikationen ergaben, bestätigen die Hypothese, dass zwischen den App Bewertungen und den Eigenschaften der Reviewer ein Zusammenhang besteht.

Ein möglicher Grund für diesen Zusammenhang könnte sein, dass die Persönlichkeitseigenschaften einer Person sich darauf auswirken, wie positiv oder negativ ihr Bewertungsverhalten ist bzw. wie positiv oder negativ diese Person Kritik ausdrückt. Auffallend ist, dass die Reviewer von positiven Bewertungen am extrovertiertesten, freundlichsten und gewissenhaftesten sind. Außerdem besitzen sie die niedrigsten Werte an Neuroticism und Openness. All diese Persönlichkeitseigenschaften unterstützen die Theorie, dass die Reviewer positiven Feedbacks freundlichen und positiven Wesens sind und daher vielleicht auch positiver bewerten. Das könnte auch daran liegen, dass diese Menschen schneller zu begeistern sind als andere. Die Verfasser der negativen Bewertungen sind den Ergebnissen der Analysen zu Folge insgesamt negativer und verschlossener, weshalb sie zu negativeren Bewertungen neigen könnten.

Da die Verfasser der positiven Reviews auch im Gegensatz zu den unzufriedenen und neutralen Usern die niedrigsten Werte an Openness besitzen, wäre es auch denkbar, dass die zufriedenen User von den Applikationen überzeugter waren, da diese weniger verschiedenen Applikationen ausprobiert haben. Die Reviewer, die besonders offen für neue Erfahrungen sind, haben möglicherweise mehr Erfahrungen mit verschiedenen Apps und besitzen deswegen mehr Vergleichsmöglichkeiten. Das wäre eine weitere Erklärung, warum diese dann schlechte oder neutrale Sterne-Bewertungen abgegeben haben.

Bei einer weiteren Theorie wird davon ausgegangen, dass die User, die positivere Bewertungen vergeben, mit den Applikationen zufriedener sind, weil diese auf User mit diesen Persönlichkeitseigenschaften ausgelegt sind. Es könnte sein, dass die Applikationen, deren Daten für diese Studie verwendet wurden, insgesamt mehr auf

den Persönlichkeitstyp dieser Menschen angepasst waren und ihnen deswegen besser gefallen haben. Da die Analysen auch für jede Applikation separat durchgeführt wurden, konnten folgende Erkenntnisse gewonnen werden: Bei den Persönlichkeitseigenschaften Conscientiousness, Extraversion und Neuroticism existieren immer exakt die selben Populationsverhältnisse wie in Tabelle A.7, obwohl die Applikationen teilweise recht verschieden sind (z.B. im Verwendungszweck). Diese drei Persönlichkeitseigenschaften scheinen daher einen größeren Einfluss auf das allgemeine Bewertungsverhalten der User zu haben und weniger von spezifischen Applikationen abhängig zu sein. Bei diesen Persönlichkeitseigenschaften bestehen außerdem die größten Unterschiede zwischen den Reviewern der verschiedenen Sterne-Bewertungen. Die anderen zwei Eigenschaften sind möglicherweise stärker dafür verantwortlich, ob eine Person mit einer spezifischen App zufrieden ist und scheinen sich weniger auf das allgemeine Bewertungsverhalten auszuwirken.

6.4. Positive Korrelation zwischen Sterne-Bewertungen und Sentiments

Die Ergebnisse der vierten Forschungsfrage bestätigen die Hypothese, dass zwischen der Sterne Bewertung der User und deren Stimmung eine positive Korrelation besteht. Das bedeutet, je höher die Stimmung der User, desto höher die Zufriedenheit mit der App (Sterne-Bewertung). Daher kann geschlossen werden, dass die Sentiments, die die User in ihren Reviews ausdrückten, mit der Zufriedenheit der User bezüglich der App gleichgesetzt werden können und nicht auf Grund von exogenen Einflüssen entstanden sind.

Bei der Analyse der gesamten und positiven Sentiments wurden sowohl für den gesamten Datensatz, als auch für jede Applikation separat ausschließlich signifikante, positive Werte identifiziert. Bei den negativen Stimmungswerten wurden jedoch nur sehr niedrige positive Korrelationen festgestellt. Bei der Trennung nach Applikationsdaten konnten nur drei von acht Korrelationen der negativen Sentiments als signifikant nachgewiesen werden. Der Grund hierfür könnte mit der Berechnung der Stimmungswerte zusammenhängen. Für die Stimmung des Users werden nur entweder die positiven oder negativen Emotionen berücksichtigt. Drückt ein Kommentar eine Vermischung beider Emotionen aus, wird trotzdem davon ausgegangen, dass der Review entweder nur negativ oder positiv ist. Das ist eine große Limitation dieses Ansatzes und in zukünftigen Arbeiten könnte versucht werden die beiden Emotionen miteinander zu verrechnen. Es könnte auch überprüft werden, ob man eine Verbesserung der Ergebnisse erzielen kann, wenn ein anderer Faktor als 1.5 bei der Gewichtung der negativen Emotionen verwendet wird. Denn die negativen Sentiments scheinen nur eine geringe Aussagekraft über die Sterne Vergabe zu besitzen. Dieser Theorie zu Folge müssten die Applikationen mit den höchsten Korrelationen, also Tripadvisor und WhatsApp, die geringste Vermischung an Gefühlen besitzen. Eine andere Erklärung wäre, dass vor allem diese Apps sehr wenige negative Sentiments besitzen, die die Korrelationen mit

ihrer geringen Aussagekraft verringern könnten.

6.5. Abhängigkeit der Sentiments von der Versionsnummer

Die Hypothese, dass die Sentiments eines Kommentars abhängig von der Versionsnummer sind, konnte beim analysierten Datensatz bestätigt werden. Durch solche Untersuchungen können die Hersteller einer App feststellen, ob und wie sich die Sentiments während der Softwareevolution verändert haben. Dadurch kann beispielsweise ermittelt werden, wie sich die Zufriedenheit der User von Version zu Version verändert hat und darauf entsprechend reagiert werden.

7. Conclusion

A. Appendix

Ergebnisse Frage 1

Application	Neuroticism	Openness	Agreeableness	Extraversion	Conscien.
Angry Birds	3.34 [4]	6.20 [1]	-6.96 [1]	8.00 [4]	-6.66 [1]
Dropbox	2.44 [3]	9.09 [2]	-5.41 [3]	3.33 [2]	-4.00 [2]
Evernote	2.85 [3]	9.68 [2]	-5.41 [3]	5.27 [3]	-2.70 [3]
Adobe Reader	2.66 [3]	11.22 [3,4]	-5.87 [2,3]	2.90 [1]	-4.05 [2]
Tripadvisor	0.00 [2]	13.16 [6]	-6.25 [2]	8.34 [4]	-2.56 [4]
Booking	-4.40 [1]	12.90 [4,5]	-5.88 [2]	10.72 [5]	0.00 [6]
Pinterest	0.00 [2]	4.76 [1]	-4.76 [4]	11.42 [6]	-2.13 [5]
WhatsApp	0.00 [2]	10.00 [3]	-4.55 [5]	9.67 [5]	-2.94 [3]

Table A.1.: Gesamter Datensatz: Ergebnisse Frage 1

Application	**Neuroticism**	**Openness**	**Agreeableness**	**Extraversion**	**Conscien.**
Angry Birds	0.00 [3]	9.10 [1,2]	-7.14 [1]	10.63 [3]	-5.97 [1]
Dropbox	0.00 [3]	12.32 [2,3]	-5.33 [2,3]	3.89 [1]	-3.95 [2]
Evernote	1.72 [3]	12.49 [3]	-5.88 [2,3]	4.76 [2]	-2.94 [3]
Adobe Reader	2.36 [3]	14.02 [3,4]	-6.51 [1,2,3]	2.98 [1]	-4.35 [1,2]
Tripadvisor	0.00 [2]	14.28 [4]	-6.98 [1,2]	6.38 [2]	-3.7 [2]
Booking	-5.16 [1]	14.29 [3,4]	-5.71 [1,2,3]	10.26 [3]	0.00 [4]
Pinterest	-4.00 [1]	8.81 [1]	-4.55 [3]	13.08 [4]	0.00 [4]
WhatsApp	-1.75 [2]	9.09 [1]	-4.55 [3]	10.00 [3]	-3.45 [2,3]

Table A.2.: Großbritannien: Ergebnisse Frage 1

Application	Neuroticism	Openness	Agreeableness	Extraversion	Conscien.
Angry Birds	3.42 [4]	5.95 [1]	-6.90 [1]	8.00 [4]	-6.67 [1]
Dropbox	1.67 [3,4]	11.73 [3,4]	-5.20 [2,3]	2.75 [1]	-4.26 [2]
Evernote	0.96 [3]	12.00 [3]	-5.32 [2,3]	5.67 [2,3]	-2.13 [3,4]
Adobe Reader	2.35 [2,3,4]	12.61 [3,4]	-6.15 [1,2,3]	3.23 [1,2]	-3.41 [2,3]
Tripadvisor	0.00 [2]	15.15 [4]	-6.25 [1,2]	7.14 [3,4]	-2.46 [3,4]
Booking	-4.00 [1]	12.00 [2,3,4]	-4.76 [1,2,3]	9.52 [4,5]	-1.43 [3,4]
Pinterest	-2.08 [1,2]	6.91 [1,2]	-4.76 [3]	11.11 [5]	-2.27 [4]
WhatsApp	-1.82 [1,2]	9.37 [3]	-5.00 [3]	10.00 [5]	-2.86 [3]

Table A.3.: Kanada: Ergebnisse Frage 1

Application	Neuroticism	Agreeableness	Extraversion	Conscien.
Angry Birds	-9.53 [1,2]	-4.76 [1]	23.81 [1]	-4.76 [1,2]
Dropbox	-1.78 [1,2]	-5.36 [1]	4.76 [1,2]	-3.23 [1,2]
Evernote	0.00 [1,2]	-5.81 [1]	7.46 [1]	-2.41 [1,2]
Adobe Reader	0.00 [1,2]	-4.35 [1]	3.03 [1,2]	-3.23 [1,2]
Tripadvisor	-2.70 [1,2]	-6.45 [1]	7.69 [1,2]	0.00 [1,2]
Booking	-6.06 [1,2]	-5.88 [1]	8.82 [1,2]	-2.94 [1,2]
Pinterest	-3.51 [1]	-5.34 [1]	11.22 [2]	0.00 [2]
WhatsApp	0.00 [2]	-4.76 [1]	7.38 [1]	-3.85 [1]

Table A.4.: Singapur: Ergebnisse Frage 1

Application	Neuroticism	Openness	Agreeableness	Extraversion	Conscien.
Angry Birds	3.42 [3]	5.95 [1]	-6.90 [1]	8.00 [1,3]	-6.67 [1]
Dropbox	2.63 [3]	8.57 [2]	-5.41 [3,4]	3.23 [1]	-4.00 [2]
Evernote	2.99 [3]	9.36 [2]	-5.41 [4]	5.33 [2]	-2.70 [3]
Adobe Reader	2.74 [3]	10.71 [2,3,4]	-5.83 [2,3,4]	2.78 [1]	-4.10 [2]
Tripadvisor	-0.88 [2]	12.77 [4]	-6.06 [2]	9.00 [4]	-2.33 [4]
Booking	-4.17 [1]	12.30 [3,4]	-6.08 [1,2,3]	11.11 [5,6]	0.00 [6]
Pinterest	0.00 [2]	4.65 [1]	-4.76 [5]	11.21 [6]	-2.17 [5]
WhatsApp	0.00 [2]	10.32 [3]	-4.55 [6]	9.80 [5]	-2.63 [4]

Table A.5.: Vereinigte Staten: Ergebnisse Frage 1

Application	Neuroticism	Openness	Agreeableness	Extraversion	Conscien.
Angry Birds	3.39 [3]	6.00 [1]	-6.90 [1]	8.00 [3]	-6.67 [1]
Dropbox	2.63 [3]	8.57 [2]	-5.41 [3,4]	3.23 [1]	-4.00 [2]
Evernote	2.98 [3]	9.37 [2]	-5.41 [3,4]	5.29 [2]	-2.70 [3]
Adobe Reader	2.86 [3]	10.72 [2,3,4]	-5.77 [2,3,4]	2.71 [1]	-4.17 [2]
Tripadvisor	-0.28 [2]	12.72 [4]	-6.06 [2]	8.82 [3]	-2.28 [4]
Booking	-4.17 [1]	12.12 [3,4]	-6.08 [1,2,3]	11.11 [4,5]	0.00 [6]
Pinterest	0.00 [2]	4.62 [1]	-4.76 [5]	11.36 [5]	-2.15 [5]
WhatsApp	0.00 [2]	10.23 [3]	-4.55 [6]	9.89 [4]	-2.63 [4]

Table A.6.: Afrika-Länder: Ergebnisse Frage 1

Ergebnisse Frage 2

Land	Openness	Agreeableness	Extraversion	Conscien.
Vereinigte Staaten	6.39 [1]	-4.76 [2]	9.86 [3]	-2.38 [3]
Afrika-Länder	6.35 [1]	-4.76 [2]	10.00 [3]	-2.38 [3]
Großbritannien	10.00 [3]	-4.76 [2]	9.52 [2]	-2.86 [2]
Kanada	8.00 [2]	-5.16 [1]	9.30 [2]	-2.94 [2]
Singapur	12.00 [4]	-4.76 [1,2]	7.50 [1]	-3.64 [1]

Table A.7.: Gesamter Datensatz: Ergebnisse Frage 2

Land	Neuroticism
Vereinigte Staaten	3.42 [2]
Afrika-Länder	3.39 [2]
Großbritannien	0.00 [1]
Kanada	3.42 [2]
Singapur	-9.53 [1,2]

Table A.8.: Angry Birds: Ergebnisse Frage 2

Land	Neuroticism	Openness
Vereinigte Staaten	2.63 [2]	8.57 [1]
Afrika-Länder	2.63 [2]	8.57 [1]
Großbritannien	0.00 [1]	12.32 [2]
Kanada	1.67 [1,2]	11.73 [1,2]
Singapur	-1.78 [1,2]	15.39 [2]

Table A.9.: Dropbox: Ergebnisse Frage 2

Land	Neuroticism	Openness	Agreeableness	Conscien.
Vereinigte Staaten	2.99 [2]	9.36 [1]	-5.41 [2]	2.99 [2]
Afrika-Länder	2.98 [2]	9.37 [1]	-5.41 [2]	2.98 [2]
Großbritannien	1.72 [1]	12.49 [2]	-5.88 [1]	1.72 [2]
Kanada	0.96 [1]	12.00 [2]	-5.31 [1,2]	0.96 [1]
Singapur	0.00 [1,2]	9.57 [1,2]	-5.81 [1,2]	0.00 [1,2]

Table A.10.: Evernote: Ergebnisse Frage 2

Land	Openness	Agreeableness	Extraversion	Conscien.
Vereinigte Staaten	12.77 [1]	-6.06 [2]	9.00 [2]	-0.88 [1]
Afrika-Länder	12.72 [1]	-6.06 [2]	8.82 [2]	-0.28 [1]
Großbritannien	14.28 [2]	-6.98 [1]	6.38 [1]	0.00 [2]
Kanada	15.15 [2]	-6.25 [1,2]	7.14 [1,2]	0.00 [1]
Singapur	19.05 [2]	-6.45 [1,2]	7.69 [1,2]	-2.70 [1,2]

Table A.11.: Tripadvisor: Ergebnisse Frage 2

Land	Neuroticism	Openness	Agreeableness	Extraversion	Conscien.
Vereinigte Staaten	0.00 [3]	4.65 [1]	-4.76 [1]	11.21 [1]	0.00 [2]
Afrika-Länder	0.00 [3]	4.62 [1]	-4.76 [1]	11.36 [1]	0.00 [2]
Großbritannien	-4.00 [1]	8.81 [3]	-4.55 [2]	13.08 [2]	-4.00 [1]
Kanada	-2.08 [2]	6.91 [2]	-4.76 [1]	11.11 [1]	-2.08 [2]
Singapur	-3.51 [1,2,3]	9.68 [2,3]	-5.34 [1,2]	11.21 [1,2]	-3.51 [1,2]

Table A.12.: Pinterest: Ergebnisse Frage 2

Land	Neuroticism	Openness	Extraversion	Conscien.
Vereinigte Staaten	0.00 [2]	10.32 [2]	9.80 [2]	0.00 [1]
Afrika-Länder	0.00 [2]	10.23 [2]	9.89 [2]	0.00 [1]
Großbritannien	-1.75 [1]	9.09 [1]	10.00 [2]	-1.75 [2]
Kanada	-1.82 [1,2]	9.37 [1,2]	10.00 [2]	-1.82 [1]
Singapur	0.00 [3]	11.83 [3]	7.38 [1]	0.00 [3]

Table A.13.: WhatsApp: Ergebnisse Frage 2

Ergebnisse Frage 3

Land	Neuroticism	Openness	Agreeableness	Extraversion	Conscien.
Positiv	-2.27 [1]	5.76 [1]	-4.55 [3]	12.50 [3]	0.00 [3]
Neutral	2.13 [2]	8.70 [2]	-6.41 [1]	4.76 [2]	-3.85 [2]
Negativ	4.54 [3]	9.09 [2]	-5.88 [2]	2.52 [1]	-5.56 [1]

Table A.14.: Gesamter Datensatz: Ergebnisse Frage 3

Land	Neuroticism	Openness	Agreeableness	Extraversion	Conscien.
Positiv	0.00 [1]	4.35 [1]	-6.52 [2]	13.33 [3]	-4.35 [3]
Neutral	3.13 [2]	5.13 [1,2]	-6.78 [1]	6.38 [2]	-5.77 [2]
Negativ	4.84 [3]	7.27 [2]	-7.14 [1,2]	5.71 [1]	-8.34 [1]

Table A.15.: Angry Birds: Ergebnisse Frage 3

Land	Neuroticism	Openness	Agreeableness	Extraversion	Conscien.
Positiv	0.00 [1]	8.47 [1]	-5.13 [2]	7.69 [3]	-2.47 [3]
Neutral	1.85 [2]	11.40 [2]	-6.12 [1]	3.55 [2]	-3.85 [2]
Negativ	4.21 [3]	8.95 [1]	-5.17 [2]	0.00 [1]	-5.40 [1]

Table A.16.: Dropbox: Ergebnisse Frage 3

Land	Neuroticism	Openness	Agreeableness	Extraversion	Conscien.
Positiv	1.05 [1]	8.48 [1]	-4.92 [3]	8.00 [3]	-0.96 [3]
Neutral	2.43 [2]	11.90 [3]	-6.90 [1]	3.08 [2]	-3.45 [2]
Negativ	5.12 [3]	11.39 [2]	-5.88 [2]	1.32 [1]	-5.26 [1]

Table A.17.: Evernote: Ergebnisse Frage 3

Land	Neuroticism	Extraversion	Conscien.
Positiv	0.00 [1]	6.45 [3]	-2.08 [3]
Neutral	1.89 [2]	3.33 [2]	-3.45 [2]
Negativ	3.85 [3]	0.00 [1]	-5.56 [1]

Table A.18.: Adobe Reader: Ergebnisse Frage 3

Land	Neuroticism	Openness	Agreeableness	Extraversion	Conscien.
Positiv	-3.28 [1]	12.90 [1]	-5.88 [2]	11.12 [3]	0.00 [3]
Neutral	0.00 [2]	12.80 [1,2]	-6.67 [1]	4.17 [2]	-4.08 [2]
Negativ	3.39 [3]	14.29 [2]	-6.90 [1]	1.00 [1]	-5.98 [1]

Table A.19.: Tripadvisor: Ergebnisse Frage 3

Land	Neuroticism	Extraversion	Conscien.
Positiv	-5.13 [1]	12.00 [3]	0.00 [3]
Neutral	2.38 [2]	7.14 [2]	-3.23 [2]
Negativ	2.30 [2]	1.95 [1]	-4.54 [1]

Table A.20.: Booking: Ergebnisse Frage 3

Land	Neuroticism	Openness	Agreeableness	Extraversion	Conscien.
Positiv	-2.50 [1]	4.44 [1]	-4.55 [3]	13.05 [3]	0.00 [3]
Neutral	2.70 [2]	7.36 [2]	-6.45 [1]	4.92 [2]	-3.85 [2]
Negativ	5.00 [3]	7.50 [2]	-6.06 [2]	2.63 [1]	-5.41 [1]

Table A.21.: Pinterest: Ergebnisse Frage 3

Land	Neuroticism	Openness	Agreeableness	Extraversion	Conscien.
Positiv	-3.42 [1]	9.68 [1]	-4.55 [2]	13.63 [3]	-1.48 [3]
Neutral	0.00 [2]	10.81 [2]	-4.92 [1]	6.46 [2]	-3.64 [2]
Negativ	3.33 [3]	9.75 [1]	-5.00 [1]	4.16 [1]	-5.13 [1]

Table A.22.: WhatsApp: Ergebnisse Frage 3

Ergebnisse Frage 4

	Angrybirds	Dropbox	Evernote	Adobe Reader
Sentiments	**0.4368**	**0.4506**	**0.4696**	**0.4495**
Negative Sentiments	0.0391	-0.0222	**0.1880**	0.0298
Positive Sentiments	**0.4320**	**0.4363**	**0.2433**	**0.4159**

Table A.23.: Ergebnisse Frage 4: Sentiments Korrelationen Teil 1

	Tripadvisor	Booking	Pinterest	WhatsApp
Sentiments	**0.4904**	**0.3828**	**0.4251**	**0.4871**
Negative Sentiments	**0.1444**	-0.0124	0.0183	**0.0976**
Pos. Sent.	**0.3162**	**0.2876**	**0.3313**	**0.4076**

Table A.24.: Ergebnisse Frage 4: Sentiments Korrelationen Teil 2

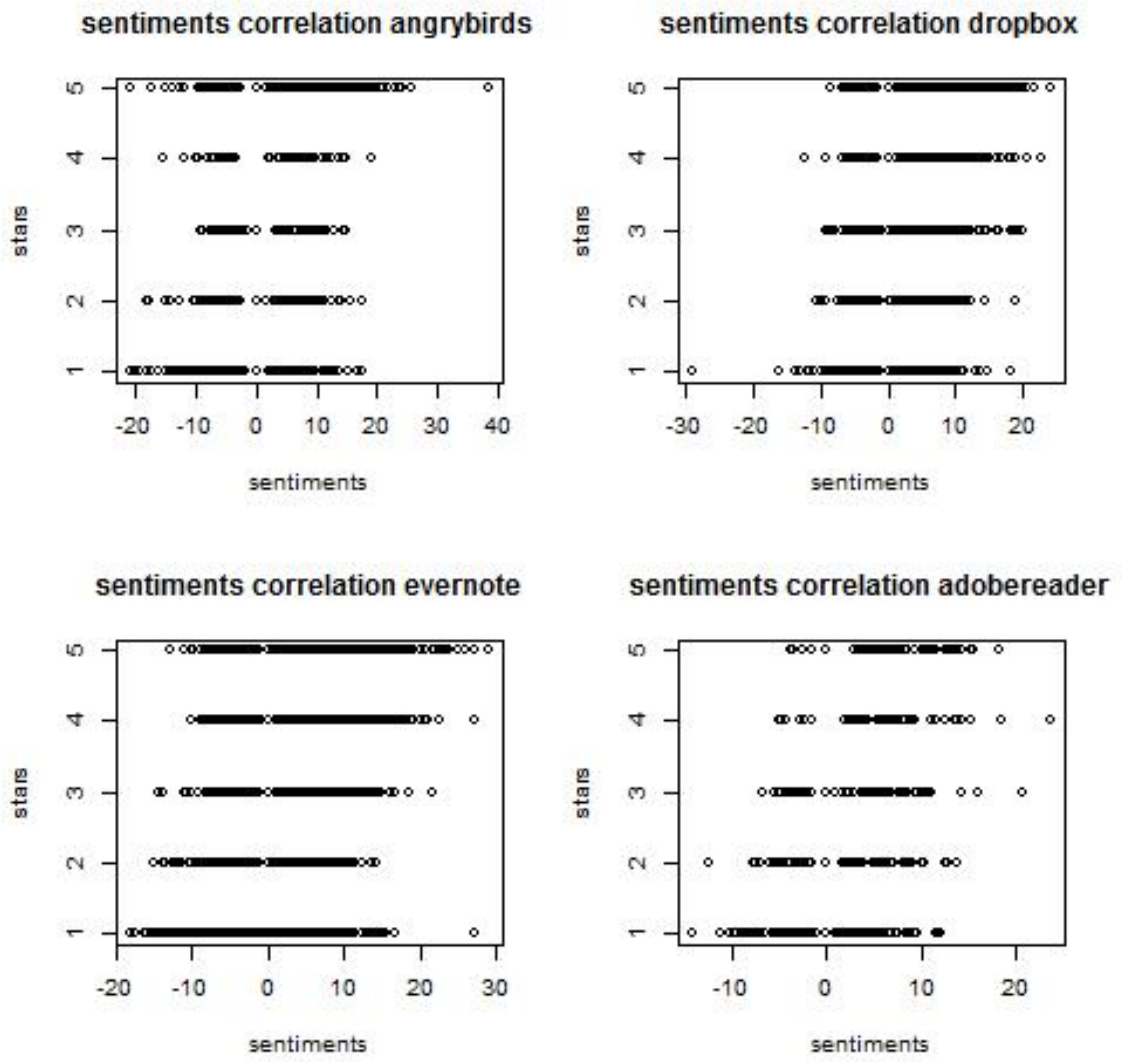


Figure A.1.: Ergebnisse Frage 4: Korrelation der Sentiments für alle Applikationen Teil 1

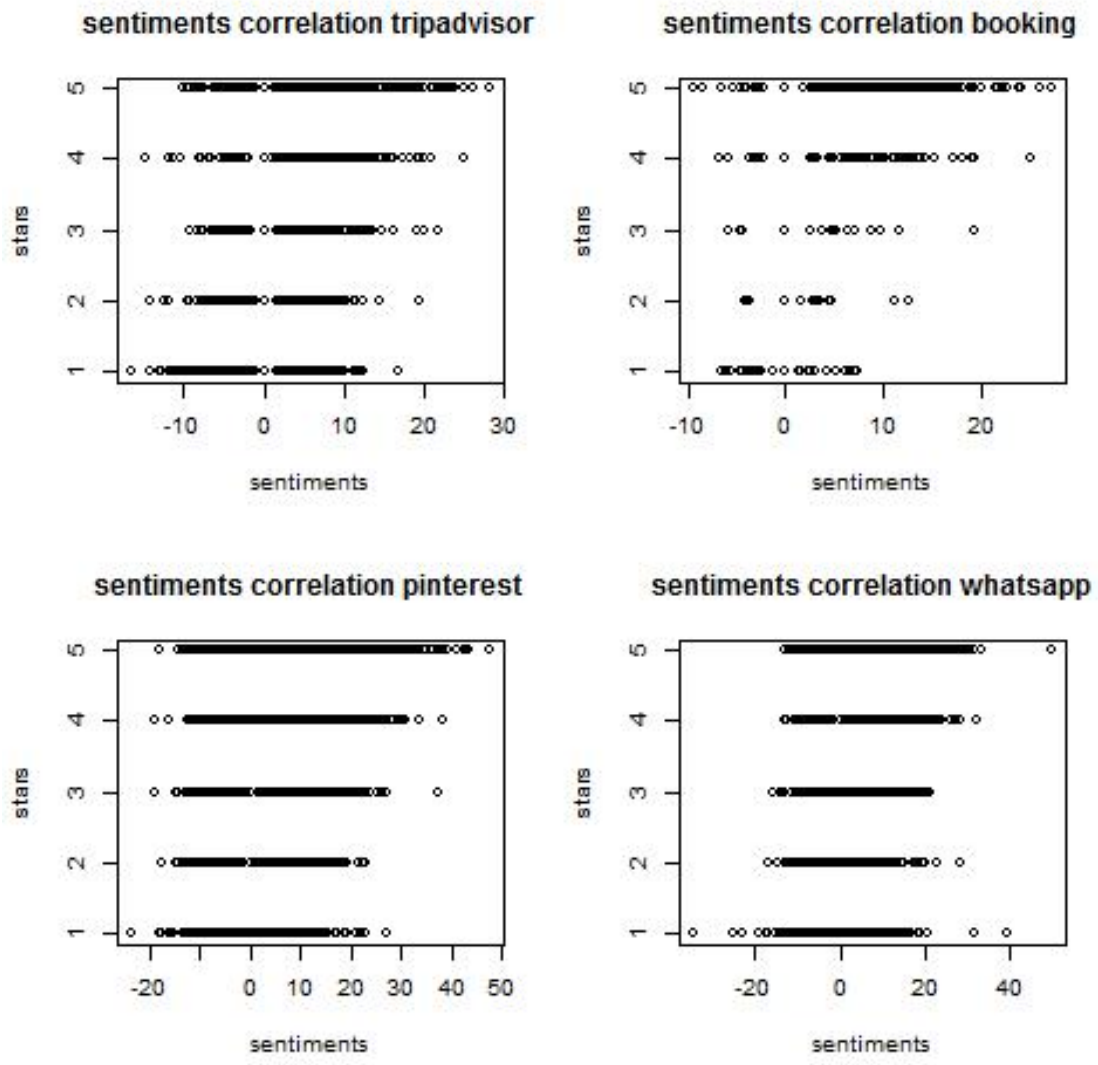


Figure A.2.: Ergebnisse Frage 4: Korrelation der Sentiments für alle Applikationen Teil 2

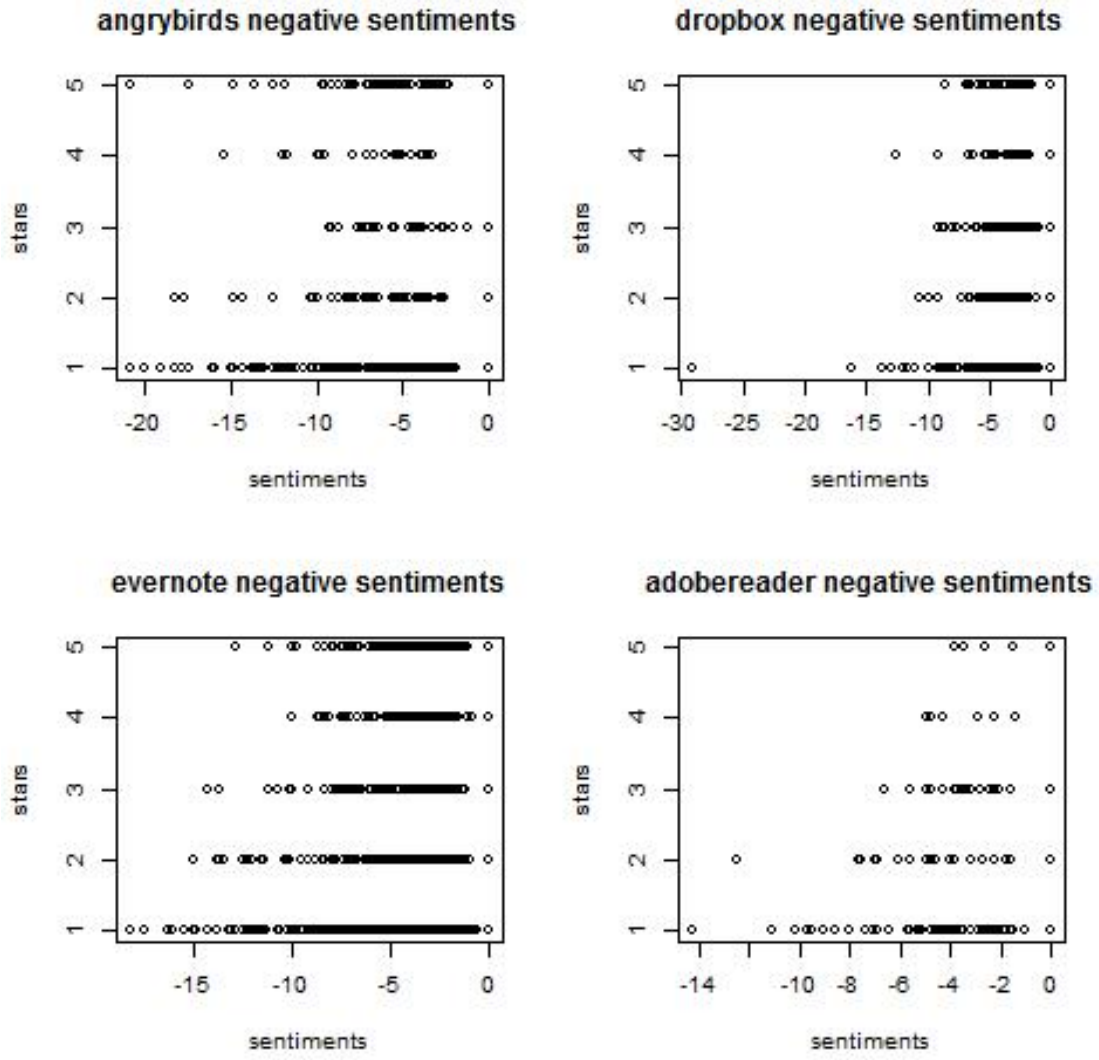


Figure A.3.: Ergebnisse Frage 4: Korrelation der negativen Sentiments für alle Applikationen Teil 1

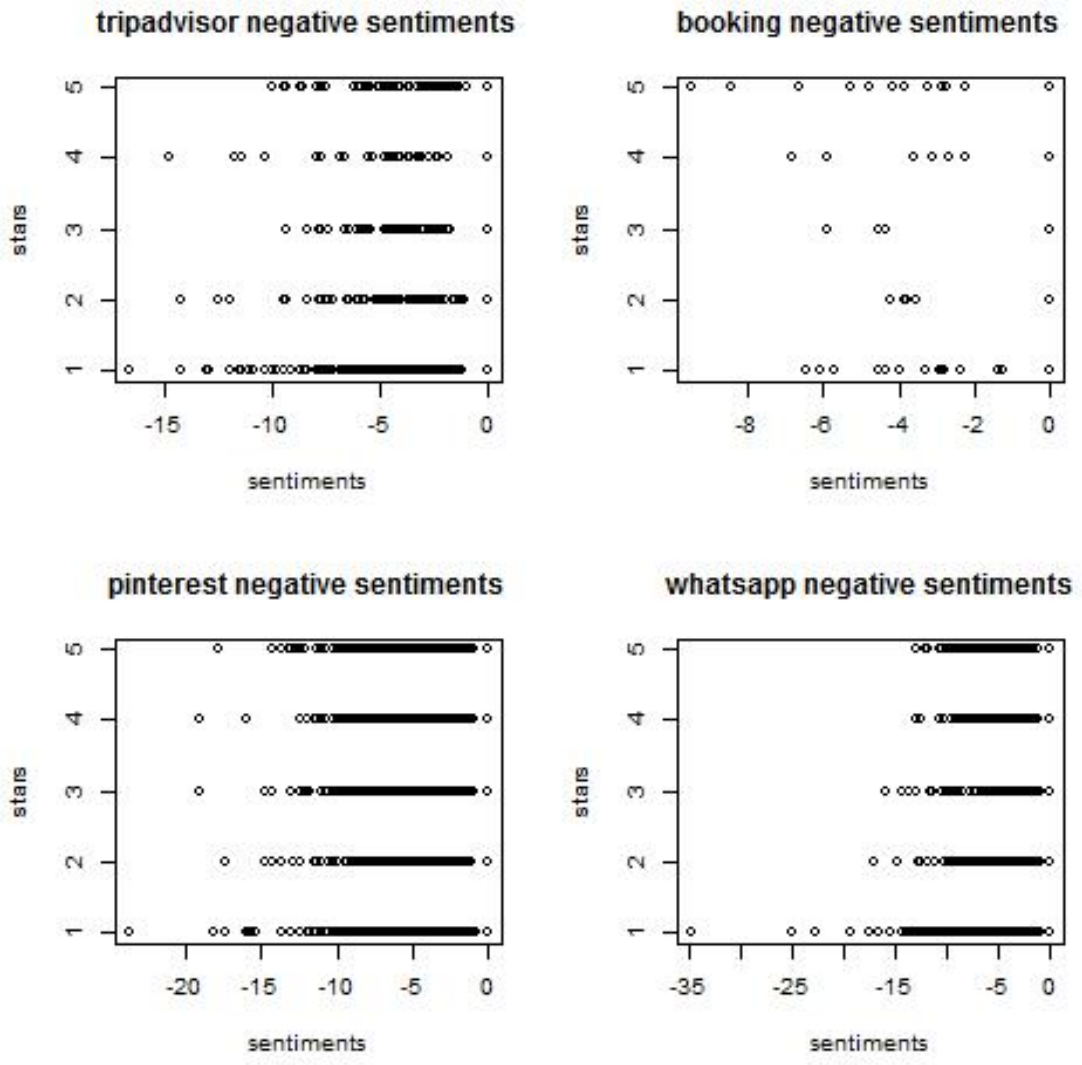


Figure A.4.: Ergebnisse Frage 4: Korrelation der negativen Sentiments für alle Applikationen Teil 2

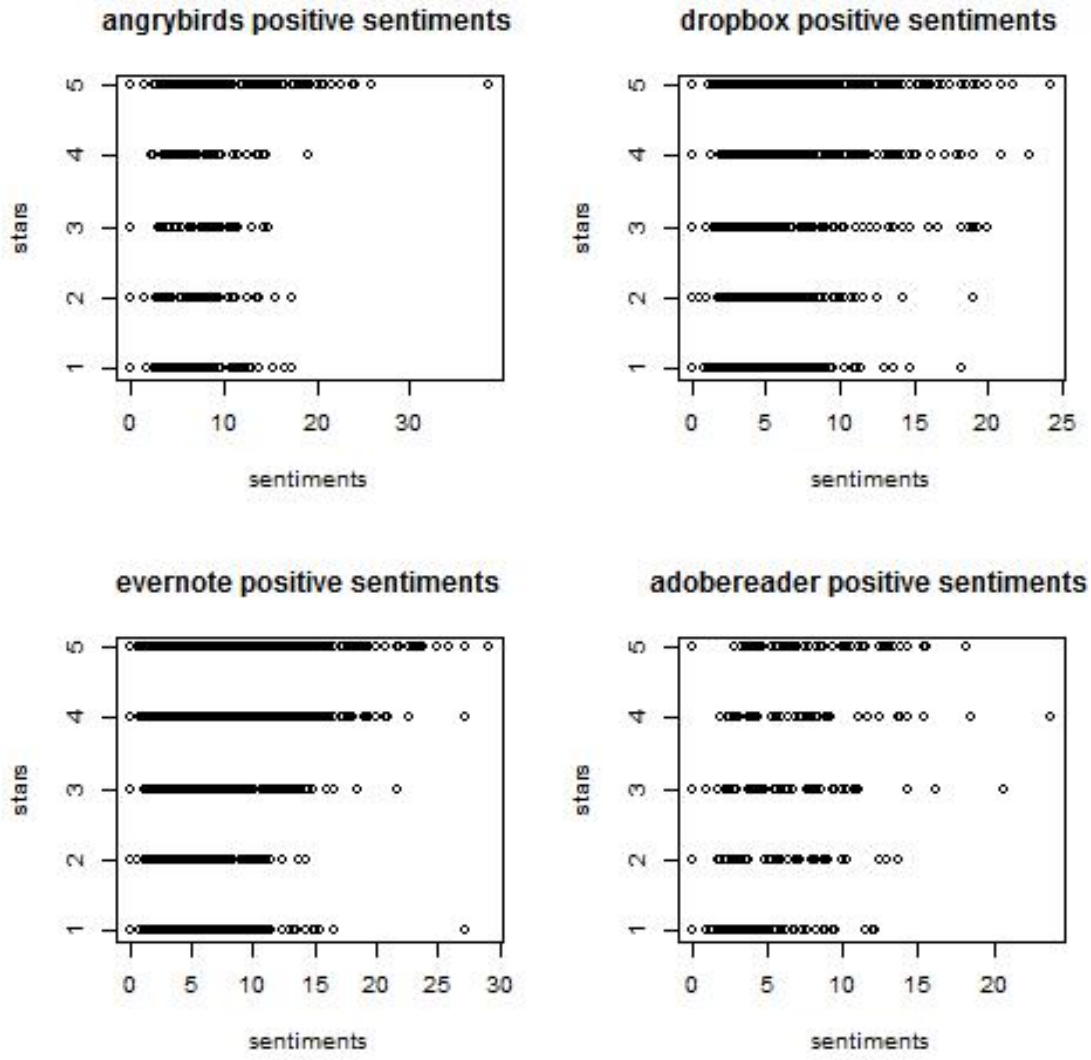


Figure A.5.: Ergebnisse Frage 4: Korrelation der positiven Sentiments für alle Applikationen Teil 1

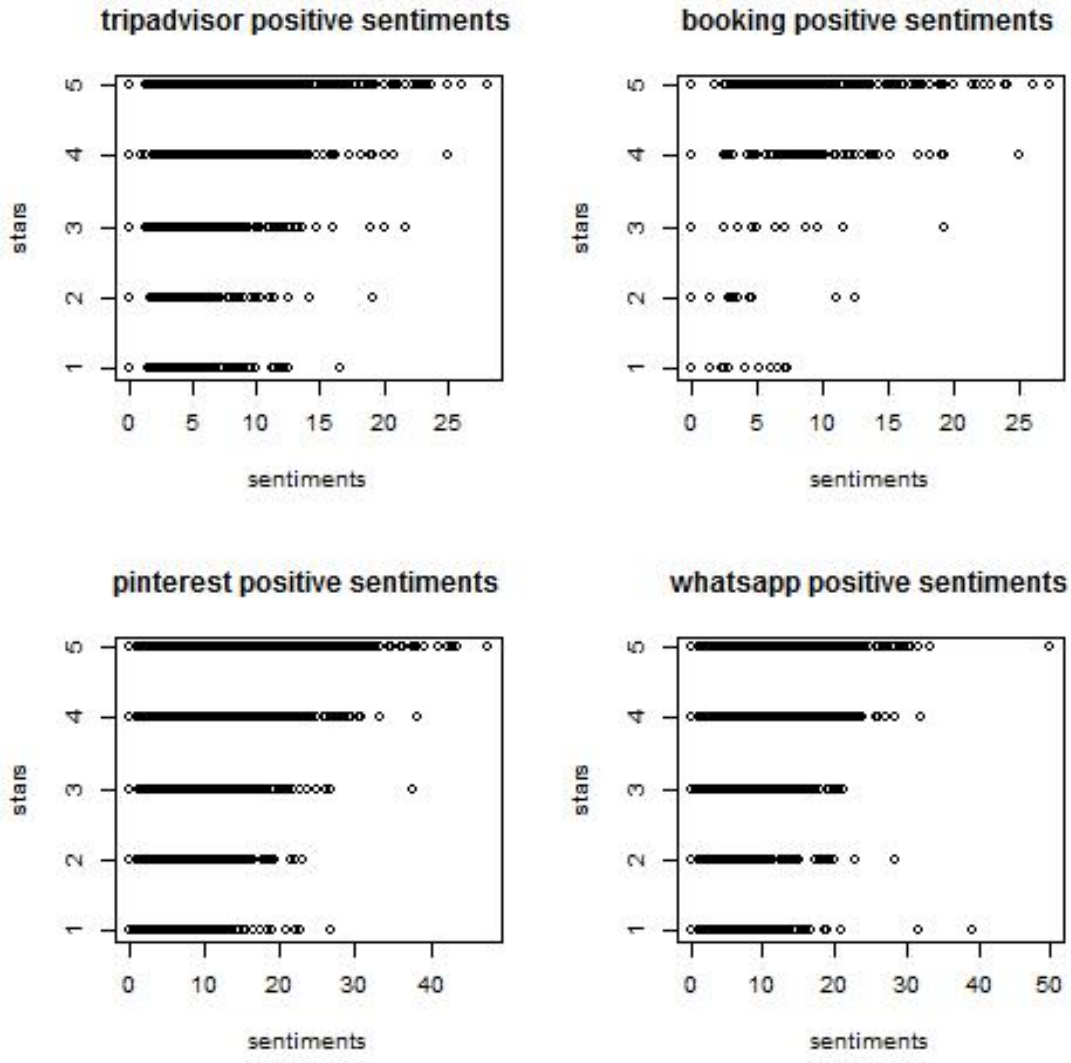


Figure A.6.: Ergebnisse Frage 4: Korrelation der positiven Sentiments für alle Applikationen Teil 2

Ergebnisse Frage 5

	4.7.1	5.1	5.2	5.3	5.4	5.4.1	5.5	5.6
5.1	1.00	-	-	-	-	-	-	-
5.2	1.00	1.00	-	-	-	-	-	-
5.3	1.00	1.00	1.00	-	-	-	-	-
5.4	1.00	1.00	1.00	1.00	-	-	-	-
5.4.1	1.00	1.00	1.00	1.00	1.00	-	-	-
5.5	1.00	1.00	1.00	1.00	1.00	1.00	-	-
5.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	-
5.6.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5.6.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5.7	1.00	1.00	1.00	0.092	1.00	0.543	1.00	1.00
5.7.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
5.8	1.00	1.00	1.00	**0.011**	1.00	0.649	1.00	1.00
5.9	1.00	1.00	1.00	0.251	1.00	1.00	1.00	1.00
5.9.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6.0.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table A.25.: Booking: Ergebnisse Frage 5 Teil 1

	5.6.1	5.6.2	5.7	5.7.1	5.8	5.9	5.9.1	6.0
5.1	-	-	-	-	-	-	-	-
5.2	-	-	-	-	-	-	-	-
5.3	-	-	-	-	-	-	-	-
5.4	-	-	-	-	-	-	-	-
5.4.1	-	-	-	-	-	-	-	-
5.5	-	-	-	-	-	-	-	-
5.6	-	-	-	-	-	-	-	-
5.6.1	-	-	-	-	-	-	-	-
5.6.2	1.00	-	-	-	-	-	-	-
5.7	1.00	1.00	-	-	-	-	-	-
5.7.1	1.00	1.00	0.351	-	-	-	-	-
5.8	1.00	0.436	1.00	0.068	-	-	-	-
5.9	1.00	1.00	1.00	1.00	1.00	-	-	-
5.9.1	1.00	1.00	1.00	1.00	1.00	1.00	-	-
6.0	1.00	1.00	1.00	1.00	0.319	1.00	1.00	-
6.0.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table A.26.: Booking: Ergebnisse Frage 5 Teil 2

Versionsnummer	4.7.1	5.1	5.2	5.3	5.4	5.4.1	5.5	5.6	
Median der Sentiments	4.760	8.330	8.960	11.370	8.110	9.380	8.000	9.090	
Versionsnummer	5.6.1	5.6.2	5.7	5.7.1	5.8	5.9	5.9.1	6.0	6.0.1
Median der Sentiments	8.330	9.680	6.450	8.700	5.810	7.320	8.330	10.340	7.655

Table A.27.: Booking: Mediane der Sentiments Für alle Versionsnummern Frage 5

	1.5.0	1.5.3	1.6.3	2.0.0	2.0.2	2.1.0	2.2.0	2.3.0
1.5.3	1.00	-	-	-	-	-	-	-
1.6.3	1.00	1.00	-	-	-	-	-	-
2.0.0	1.00	1.00	1.00	-	-	-	-	-
2.0.2	1.00	1.00	1.00	1.00	-	-	-	-
2.1.0	1.00	1.00	1.00	1.00	1.00	-	-	-
2.2.0	1.00	1.00	1.00	1.00	1.00	1.00	-	-
2.3.0	1.00	0.48492	1.00	1.00	1.00	1.00	1.00	-
3.0.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3.1.0	1.00	0.59831	1.00	1.00	1.00	1.00	1.00	1.00
3.1.2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3.2.0	1.00	1.00	1.00	1.00	1.00	1.00	**0.02771**	0.42022
3.3.0	1.00	1.00	1.00	1.00	1.00	1.00	0.80606	1.00
3.3.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3.4.0	1.00	1.00	0.67989	1.00	1.00	1.00	0.07940	1.00
3.4.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4.0.0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table A.28.: Angry Birds: Ergebnisse Frage 5 Teil 1

	3.0.0	3.1.0	3.1.2	3.2.0	3.3.0	3.3.1	3.4.0	3.4.1
1.5.3	-	-	-	-	-	-	-	-
1.6.3	-	-	-	-	-	-	-	-
2.0.0	-	-	-	-	-	-	-	-
2.0.2	-	-	-	-	-	-	-	-
2.1.0	-	-	-	-	-	-	-	-
2.2.0	-	-	-	-	-	-	-	-
2.3.0	-	-	-	-	-	-	-	-
3.0.0	-	-	-	-	-	-	-	-
3.1.0	0.09495	-	-	-	-	-	-	-
3.1.2	1.00	**< 2e-16**	-	-	-	-	-	-
3.2.0	**0.00174**	0.19087	**3.3e-05**	-	-	-	-	-
3.3.0	1.00	1.00	1.00	1.00	-	-	-	-
3.3.1	1.00	0.62225	1.00	**1.4e-08**	1.00	-	-	-
3.4.0	1.00	1.00	1.00	1.00	1.00	1.00	-	-
3.4.1	1.00	1.00	1.00	**0.00050**	1.00	1.00	1.00	-
4.0.0	1.00	1.00	1.00	**0.00015**	1.00	1.00	1.00	1.00

Table A.29.: Angry Birds: Ergebnisse Frage 5 Teil 2

Versionsnummer	1.5.0	1.5.3	1.6.3	2.0.0	2.0.2	2.1.0	2.2.0	2.3.0	
Median der Sentiments	0.00	-8.70	4.55	11.54	11.67	5.56	8.70	0.00	
Versionsnummer	3.0.0	3.1.0	3.1.2	3.2.0	3.3.0	3.3.1	3.4.0	3.4.1	4.0.0
Median der Sentiments	0.00	5.13	6.25	-4.35	-3.23	4.35	-4.76	3.77	4.76

Table A.30.: Angry Birds: Mediane der Sentiments für alle Versionsnummern Frage 5

Glossary

.

intervallskalierte Skala Eine intervallskalierte Skala ist eine Skala, auf der die Abstände zwischen zwei Merkmalen gemessen werden können.

ordinalskalierte Skala Die Merkmale einer ordinalskalierten Skala können in eine Größenordnung gebracht werden. Die Abstände zwischen den beiden Merkmalen können jedoch nicht gemessen werden.

Requirements Engineering Das Requirements Engineering ist eine iterative Vorgehensweise. Im Zuge des Requirements Engineering wird eine System- und Anforderungsspezifikation generiert, die den Bedürfnissen aller Stakeholder gerecht wird.

Reviewer Der Verfasser eines Kommentars im App Store.

Roadmap Eine Roadmap beschreibt die Vorgehensweise, die notwendig ist, um ein bestimmtes Ergebnis zu erzielen.

Sentiment Sentiments sind die Stimmungen, die durch den User ausgedrückt werden. In dieser Studie werden die Sentiments an Hand der positiven und negativen Emotionen gemessen, die im, durch den User verfassten Text, mit LIWC gemessen werden. Jedem Kommentar wird entweder der mit (-1) multiplizierte "Negative Emotion" Wert von LIWC oder der "Positive Emotion" Wert zugewiesen.

Stakeholder Ein Stakeholder ist eine Person oder eine Gruppe von Personen, die Interesse an der Erstellung eines Systems oder dessen Ergebnis haben.

Acronyms

LIWC Linguistic Inquiry and Word Count.

List of Figures

1.1. Aktivitäten des Requirements Engineerings 4
1.2. Softwareentwicklungsprozess 4

4.1. Überblick über den Forschungsansatz 14
4.2. QQ Plot Evernote Afrika-Länder 26
4.3. Boxplots von Verteilungen verschiedener Schiefen 28

5.1. Ergebnisse Frage 5: Korrelation der Sentiments mit den Sterne-Bewertungen 37
5.2. Korrelation der positiven Sentiments 38
5.3. Korrelation der negativen Sentiments 38

A.1. Ergebnisse Frage 4: Korrelation der Sentiments für alle Applikationen Teil 1 59
A.2. Ergebnisse Frage 4: Korrelation der Sentiments für alle Applikationen Teil 2 60
A.3. Ergebnisse Frage 4: Korrelation der negativen Sentiments für alle Applikationen Teil 1 61
A.4. Ergebnisse Frage 4: Korrelation der negativen Sentiments für alle Applikationen Teil 2 62
A.5. Ergebnisse Frage 4: Korrelation der positiven Sentiments für alle Applikationen Teil 1 63
A.6. Ergebnisse Frage 4: Korrelation der positiven Sentiments für alle Applikationen Teil 2 64

List of Tables

1.1. Überblick über den Wachstum des App-Angebots und der aktiven Entwickler in den Jahren 2009-2011 [BK11] 2

4.1. Überblick über Daten im App Store 17
4.2. Überblick über Daten, die in der Analyse verwendet wurden (Jahr 2013 und > 20 Wörter) 17
4.3. Korrelation der LIWC Dimensionen mit den Big Five Personality Dimensions. * p < 0.05, ** p < 0.01, two-tailed [PK99]. 23

A.1. Gesamter Datensatz: Ergebnisse Frage 1 49
A.2. Großbritannien: Ergebnisse Frage 1 49
A.3. Kanada: Ergebnisse Frage 1 50
A.4. Singapur: Ergebnisse Frage 1 50
A.5. Vereinigte Staten: Ergebnisse Frage 1 51
A.6. Afrika-Länder: Ergebnisse Frage 1 51
A.7. Gesamter Datensatz: Ergebnisse Frage 2 52
A.8. Angry Birds: Ergebnisse Frage 2 52
A.9. Dropbox: Ergebnisse Frage 2 53
A.10. Evernote: Ergebnisse Frage 2 53
A.11. Tripadvisor: Ergebnisse Frage 2 53
A.12. Pinterest: Ergebnisse Frage 2 54
A.13. WhatsApp: Ergebnisse Frage 2 54
A.14. Gesamter Datensatz: Ergebnisse Frage 3 55
A.15. Angry Birds: Ergebnisse Frage 3 55
A.16. Dropbox: Ergebnisse Frage 3 55
A.17. Evernote: Ergebnisse Frage 3 56
A.18. Adobe Reader: Ergebnisse Frage 3 56
A.19. Tripadvisor: Ergebnisse Frage 3 56
A.20. Booking: Ergebnisse Frage 3 56
A.21. Pinterest: Ergebnisse Frage 3 56
A.22. WhatsApp: Ergebnisse Frage 3 57
A.23. Ergebnisse Frage 4: Sentiments Korrelationen Teil 1 58
A.24. Ergebnisse Frage 4: Sentiments Korrelationen Teil 2 58
A.25. Booking: Ergebnisse Frage 5 Teil 1 65
A.26. Booking: Ergebnisse Frage 5 Teil 2 66
A.27. Booking: Mediane der Sentiments Für alle Versionsnummern Frage 5 . . 66

A.28.Angry Birds: Ergebnisse Frage 5 Teil 1 67
A.29.Angry Birds: Ergebnisse Frage 5 Teil 2 67
A.30.Angry Birds: Mediane der Sentiments für alle Versionsnummern Frage 5 68

Bibliography

[BHS13] B. Bazelli, A. Hindle, and E. Stroulia. "On the Personality Traits of StackOverflow Users." In: *2013 IEEE International Conference on Software Maintenance* (Sept. 2013), pp. 460–463.

[BI] Abgerufen am 08.07.14. Business Insider: Best Apps when App Store launched. URL: http://www.businessinsider.com/the-best-iphone-apps-when-the-app-store-launched-2011-5/phonesaber-was-a-app-store-obsession-for-months-1.

[BK11] R. C. Basole and J. Karla. "Entwicklung von Mobile-Platform-Ecosystem-Strukturen und -Strategien." In: *Wirtschaftsinformatik* 53.5 (Aug. 2011), pp. 301–311. ISSN: 0937-6429.

[BM13] M. Broy and D. Méndez. *Vorlesungsskritp Requirements Engineering WS 2013/14*. München: Technische Universität München, 2013.

[CC] Abgerufen am 09.07.14. Affiliate Resources Apple App Store. URL: http://www.apple.com/itunes/affiliates/resources/documentation/linking-to-the-itunes-music-store.html.

[DC13] M. De Choudhury and S. Counts. "Understanding affect in the workplace via social media." In: *Proc. of the 2013 conference on Computer supported cooperative work - CSCW '13*. Feb. 2013, pp. 303–316.

[DM93] R. J. Dolan and J. M. Matthews. "Maximizing the Utility of Customer Product Testing: Beta Test Design and Management." In: *Journal of Product Innovation Management* 10.4 (Sept. 1993), pp. 318–330. DOI: 10.1111/1540-5885.1040318.

[Gol+90] L. R. Goldberg, O. P. John, H. Kaiser, K. Lanning, and D. Peabody. *PERSONALITY PROCESSES AND INDIVIDUAL DIFFERENCES. An Alternative " Description of Personality ": The Big-Five Factor Structure*. 1990.

[GP] J. Grudin and J. Pruitt. "Personas, Participatory Design and Product Development : An Infrastructure for Engagement." In: ().

[GRT11] J. Golbeck, C. Robles, and K. Turner. "Predicting personality with social media." In: *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11* (2011), p. 253.

[Hau12] S. Haug. *Statistik für Betriebswirtschaftslehre, Einführung mit R*. München: Technische Universität München, 2012.

[HS12] J. Hedderich and L. Sachs. *Angewandte Statistik: Methodensammlung mit R*. Springer, 2012.

[JB87] C. M. Jarque and A. K. Bera. "A Test for Normality of observation and Regression Residuals." In: (1987), pp. 163–172.

[LIWC] Abgerufen am 09.07.14. LIWC: Linguistic Inquiry and Word Count. URL: `http://www.liwc.net/descriptiontable1.php`.

[MB12] F. Matthes and B. Brügge. *Vorlesungsskritp Einführung in die Softwaretechnik SS 2012*. München: Technische Universität München, 2012.

[PBF] J. W. Pennebaker, R. J. Booth, and M. E. Francis. *Operator ' s Manual Linguistic Inquiry and Word Count : LIWC 2007*.

[Pen+] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth. *LIWC2007LanguageManual*.

[PK99] J. W. Pennebaker and L. A. King. "Linguistic Styles: Language Use as an Individual Difference." In: *Journal of Personality and Social Psychology*. 1999, pp. 1296–1312.

[Sac99] L. Sachs. *Angewandte Statistik: Anwendungen statistischer Methoden*. Springer, 1999.

[SentiStr] Abgerufen am 09.07.14. SentiStrength Java Manual. URL: `sentistrength.wlv.ac.uk/documentation/SentiStrengthJavaManual.doc`.

[Sta] Abgerufen am 08.07.14. Statista: Anzahl der Apps in Top App Stores 2013. URL: `http://de.statista.com/statistik/daten/studie/208599/umfrage/anzahl-der-apps-in-den-top-app-stores/`.

[TP10] Y. R. Tausczik and J. W. Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods." In: *Journal of Language and Social Psychology* 29.1 (2010), pp. 24–54.

[WRG13] C. Wienberg, M. Roemmele, and A. S. Gordon. "Content-based similarity measures of weblog authors." In: *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13* (2013), pp. 445–452.