

Hör auf deine Nutzer – Qualitätsverbesserung durch leichtgewichtige Analysen

Simon André Scherr, Frank Elberzhager, Selina Meyer
Fraunhofer IESE, Fraunhofer-Platz 1, 67663 Kaiserslautern
{simon.scherr, frank.elberzhager, selina.meyer}@iese.fraunhofer.de

Zusammenfassung:

Der Einfluss von Nutzern auf den Erfolg von mobilen Apps ist in den letzten Jahren deutlich gestiegen, was beispielsweise an einfachen Bewertungsmöglichkeiten in Apps Stores liegt. Entwickler müssen den Spagat aus schnell am Markt platzierten Innovationen und hoher Qualität bewältigen. Wir haben einen Qualitätssicherungsansatz entwickelt, welcher Nutzerfeedback beim Testen berücksichtigt, um Apps zu verbessern. In diesem Beitrag stellen wir unseren Analyseprozess und den Nutzen von schnellen und leichtgewichtigen Analysen vor. Herausstechende Kernelemente des Prozesses sind Analysen der im Feedback enthaltenen Emojis, die Erkennung von Trends in den Daten sowie das Ableiten von Handlungsempfehlungen.

Schlüsselworte: Qualitätssicherung, Mobile Apps, Nutzerfeedback, Emojis

1. Einleitung und Motivation

Die Welt, in der wir leben, ist heute zunehmend digital vernetzt. Einen großen Teil dieser Vernetzung machen mobile Apps aus. Diese sind zunehmend intelligenter, komplexer und häufig Teil von digitalen Ökosystemen. Anwender fordern eine fehlerfreie Nutzererfahrung und innovative Ideen. Für Entwickler bestehen neben der Sicherstellung von hoher Produktqualität weitere Herausforderungen: eine kurze Time-to-Market, regelmäßige Updates sowie geringe Budgets für Entwicklung. Nutzer neigen heute dazu, Apps schnell wieder zu löschen, wenn diese nicht ihre Erwartungen erfüllen. Aus diesem Grund müssen Qualitätsprobleme in mobilen Apps schnell erkannt und behoben werden. Wir plädieren deswegen für einen Qualitätssicherungsansatz, der Nutzer enger einbindet und ihr Feedback stärker berücksichtigt.

2. Hintergrund und Herausforderung

In [1] haben wir unsere Vision der automatischen Einbindung von Nutzerfeedback - Opti4Apps (siehe Abbildung 1) - vorgestellt. Unser Fokus lag hierbei auf theoretischen Grundlagen sowie der Klassifikation von Feedback. Erste praktische Erfahrungen mit Nutzerfeed-

back haben wir in [2] präsentiert, jedoch noch ohne einem definierten Analyseprozess zu folgen. Anspruch von Opti4Apps ist es, eine leichtgewichtige Integration von Feedback in den Entwicklungsprozess zu realisieren. Hierbei gibt es zwei Arten von Feedback. Zum einen arbeitet ein Software-Agent in der App und sammelt *Nutzungsdaten*. Zum anderen sammeln Datenkollektoren *textuelles Feedback* aus zahlreichen unterschiedlichen Datenquellen wie App Stores, sozialen Medien, Foren und Supportbereichen. Die dadurch erfassten Daten werden analysiert, was einerseits Ideen für neue Features gibt, andererseits auch Fehler aufzeigt. Ziel ist es, Unterstützung für Entwicklungsorganisationen zu bieten und hierbei vor allem Produktmanagern in Form eines Web-basierten Dashboards die Möglichkeit zu geben, fundierte Entscheidungen zur Qualitätssteigerung der eigenen Produkte zu fällen.

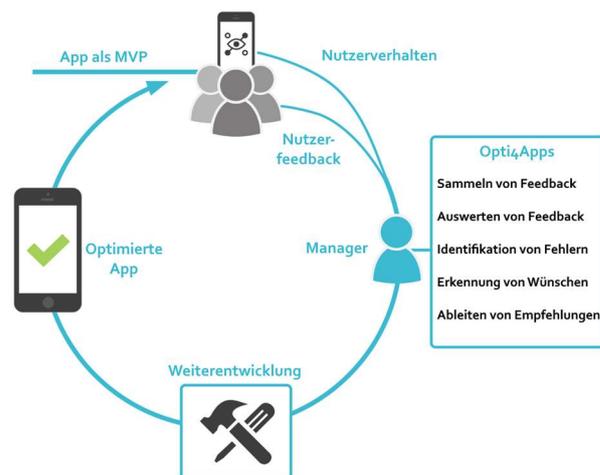


Abbildung 1 Opti4Apps im Entwicklungsprozess

Wir haben uns hierbei stark auf den Teil des textuellen Feedbacks konzentriert. Da die Entwicklung der Techniken zu „Natural Language Processing“ noch in der frühen Phase steckt, haben wir unseren Fokus auf Analysen gerichtet, die kein automatisiertes Verständnis der Sprache erfordern. Wichtig hierbei war, dass Lösungen geschaffen werden, die mit den schnellen Innovationszyklen der Produkte umgehen können und unkompliziert anwendbar sind. Um aus den zahlreichen Quellen verwertbare Ergebnisse zu erhalten, benötigt es eine Metrik, die überall anzutreffen und intuitiv verständlich

ist. Die Berücksichtigung von Bewertungssternen funktioniert nur eingeschränkt, da diese nicht überall verfügbar sind. Betrachtet man textuelles Feedback näher, stellt man fest, dass es voller Emotionen ist. Ein Weg, wie Nutzer ihre Emotionen deutlich machen, ist die Verwendung von Emojis. Daher nutzt unsere Analyse insbesondere auch Emojis.

3. Nutzung von Emojis – Vorarbeiten

Um Emojis aus Nutzerfeedback in einem Qualitätssicherungsansatz zu verwenden ist es notwendig, ein Modell zu erarbeiten, um Emojis kategorisieren zu können sowie eine Liste von Emojis zu sammeln, welche berücksichtigt werden sollen. Zum Zweck der Einordnung von Emojis in unser Modell haben wir eine Umfrage durchgeführt, auf die wir im weiteren Verlauf des Kapitels eingehen werden.

Es existieren unterschiedliche Klassifikation von Emotionen, wie die von Ekman [3], Lazarus [4] und Plutchik [5]. Für unser Vorhaben war allerdings keines dieser Modelle ausreichend: die feingranularen Modelle lassen sich nicht auf die als Piktogramm dargestellten Emojis anwenden, aber auch die größeren Modelle berücksichtigen nicht die Besonderheiten, die bei Emojis vorherrschen. Unser Emotionsmodell sieht deswegen wie folgt aus (siehe

Abbildung 2): Es gibt zwei Ebenen. Zuerst findet wie bei Feldmann [6] eine Einteilung in Sentiments statt. Es wird zwischen „positiv“, „neutral“ und „negativ“ unterschieden. Die zweite Ebene ist die Emotionsebene und fungiert als Unterkategorie des Sentiments. Sie enthält im Wesentlichen Basisemotionen, welche einfach zu verstehen und zu unterscheiden sind.

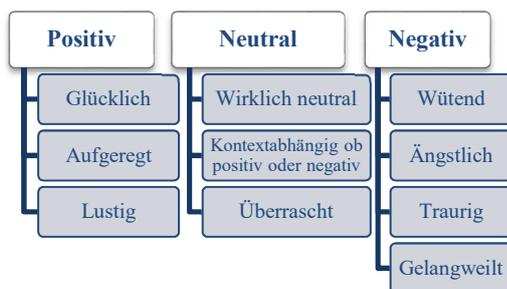


Abbildung 2 Emotionsmodell für Emojis

Für die Erfassung der Emojis nutzen wir den Unicode Standard. Der Standard enthält 2384 Emojis. Emoticons, also Folgen von Zeichen wie zum Beispiel „;-)“, stellen für uns ebenfalls Emojis dar, da sie von Emoji-Keyboards vieler Mobilgeräte auch unterstützt werden. Die Ausgangsmenge reduzierten wir auf 612 indem wir alle Emojis, die keine Emotion übermitteln (z.B. 🍌, 🚗, 🍌),

🏠, 🍷 oder Flaggen) strichen. Außerdem wurden Emojis mit gleicher Bedeutung zu Gruppen zusammengefasst, sodass unter anderem alle Kombination aus Hautfarben und Geschlechtern in einer Gruppe sind. Des Weiteren wurden gleichartige Emojis gruppiert. Jede Gruppe erhielt einen Repräsentanten, dessen Kategorisierung auf die restlichen Emojis der Gruppe übertragen werden kann. Die Einteilung wurde von drei Mitarbeitern unabhängig geprüft. Durch die Gruppierung der Emojis kamen wir auf 99 Repräsentanten.

Anschließend führten wir mit unserem Emotionsmodell und den Repräsentanten die Umfrage durch. Die Teilnehmer wurden um eine Einordnung der Emojis in unser Emotionsmodell gebeten. Hierbei sollte zuerst das Sentiment kategorisiert werden und anschließend eine entsprechende Emotion (siehe Abbildung 2). Die Fragebögen beantworteten Personen beider Geschlechts, unterschiedlichen Alters und mit unterschiedlicher Vertrautheit im Umgang mit Emojis.

Insgesamt haben wir Fragebögen von 107 Teilnehmern ausgewertet. Davon waren 53 weiblich und 54 männlich. Das Durchschnittsalter lag bei 31,41 Jahren. Die meisten Teilnehmer waren Westeuropäer oder lebten zur Zeit der Umfrage dort. 101 der Befragten gaben an, vertraut (n = 75) bis mittel vertraut (n = 26) im Umgang mit Emojis zu sein. Das Ergebnis der Umfrage ist in Tabelle 1 dargestellt. Die Auswertung¹ hat gezeigt, dass 591 der 612 Emojis auf Grund von Übereinstimmung der Befragten einem Sentiment zugewiesen werden können. Auf der Ebene der Emotionen konnten wir von den 591 Emojis 512 einer Emotion zuordnen. Insgesamt konnten wir damit zeigen, dass die Teilnehmer Emojis homogen wahrnehmen.

Tabelle 1 Kategorisierung der Emojis nach unserer Umfrage

Positiv	243	Glücklich	194
		Aufgeregt	34
		Lustig	15
Neutral	131	Wirklich neutral	4
		Kontextabhängig	97
		Überrascht	30
Negativ	217	Wütend	153
		Ängstlich	39
		Traurig	12
		Gelangweilt	13

4. Kontinuierliche Qualitätssicherung durch Trenderkennung in Feedback

Im Folgenden erläutern wir unseren im Rahmen von Opti4Apps entstandenen Qualitätssicherungsprozesses

¹ Die Liste der Emojis sowie unsere Ergebnisse sind zum verfügbar unter: <http://opti4apps.iiese.de/emoji/studyoverview.xlsx>

(Abbildung 3) zur Produktverbesserung durch die Analyse von textuellem Feedback. Zentrales Ziel ist es, die Entwicklungsorganisation möglichst wenig durch zusätzliche Aufgaben zu belasten, jedoch die Qualität der Software zu sichern.

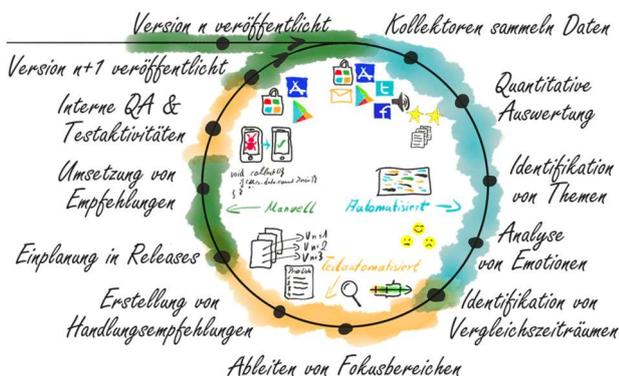


Abbildung 3 Prozess zur Qualitätsverbesserung durch Trenderkennung in Nutzerfeedback

Kernidee ist, dass das Nutzerfeedback parallel zur Entwicklung neuer Versionen der Produkte gesammelt und ausgewertet wird. Hierbei verwenden wir eine Mischung aus vollautomatisierten, teilautomatisierten und manuellen Schritten.

Kollektoren sammeln Feedback aus zahlreichen Quellen wie App Stores, soziale Medien, Foren oder Kundensupportbereichen. In der quantitativen Auswertung wird bestimmt, welches Feedback aus welcher Quelle zu welchen Zeitpunkten vorkam und ob eventuell im Feedback selbst schon Bewertungsindikatoren, wie Bewertungssterne oder Zustimmungen anderer Nutzer, vorhanden sind.

Im nächsten Schritt identifiziert der Ansatz automatisch Themen innerhalb des Feedbacks. Themen können beispielsweise einzelne Features, Fehler, Anwendungsfälle oder Screens der Anwendung sein.

Der nächste Schritt ist die Erkennung von Emotionen mittels Emojis innerhalb der Feedback-Einträge und innerhalb der zuvor erstellten Themen. Dazu nutzen wir die Ergebnisse aus den in Abschnitt 3 beschriebenen Vorarbeiten. Bis zu diesem Zeitpunkt läuft der Prozess vollständig automatisiert ab.

Danach wird das Feedback mit Blick auf frühere Produktversionen ausgewertet. Je nach Thema werden geeignete Vergleichszeiträume gewählt. Damit werden Stimmungstrends, -Wendungen sowie Veränderungen der Menge innerhalb des Feedbacks erkannt.

Basierend auf diesen Analyseschritten leitet unser Ansatz Fokusbereiche für den Produktmanager ab. Ein Fokusbereich ist eine Teilmenge des Feedbacks, welches Aufmerksamkeit bedarf. Fokusbereiche an sich sind wertneutral zu sehen, da sowohl positive als auch negative Wendungen im Feedback aufgezeigt werden.

Die anschließende Ableitung von Handlungsempfehlungen ist auf Probleme im Produkt ausgerichtet. Ziel ist es, Empfehlungen zur Produktverbesserung abzuleiten und diese zu priorisieren. Für den Fall, dass eine Handlungsempfehlung die Beseitigung eines Fehlers im Produkt ist, wird der Fehler klassifiziert und festgehalten. Damit entsteht auf lange Sicht eine Sammlung von Fehlern- und Fehlermustern zum Produkt. Ist die Handlungsempfehlung eine Abänderung des Produktes, werden neue Anforderungen abgeleitet.

Basierend auf den Handlungsempfehlungen erfolgt ein Einplanen in den Release-Plan. Daraufhin findet eine Umsetzung der Empfehlungen statt. Bei der darauffolgenden Qualitätssicherung des Produktes wird zum einen traditionelle Qualitätssicherung betrieben, aber auch ganz bewusst auf die abgeleiteten Handlungsempfehlungen geachtet. Zusätzlich werden die bereits gesammelten Fehlermuster geprüft, um so das Wiederauftreten von alten Fehlern in neuen Versionen zu vermeiden. Wenn die Version den gewünschten Qualitätsstandards entspricht, wird sie veröffentlicht.

5. Evaluation

Wir zeigen im Folgenden zwei initiale Evaluationen für unseren Prozess. Zum einen haben wir im Apple App Store die Analyse von Emojis mit der Analyse von Sternbewertungen korreliert; zum anderen haben wir anhand von verschiedenen Apps beispielhafte Auswertungen mit unserem Prozess durchgeführt. Damit liegt der Schwerpunkt der Evaluation auf der ersten Hälfte des Prozesses bis hin zum Ableiten von Fokusbereichen.

5.1 Vergleich von Emojis und Bewertungssternen

Um herauszufinden, wie sehr sich Feedback, das mithilfe von Emojis analysiert wird, mit Feedback, das unter der Berücksichtigung von Bewertungssternen analysiert wird, gleicht, haben wir 178.848 Feedback-Einträge zu diversen Apps aus dem Apple App Store gesammelt. Wir haben die darin enthaltenen Emojis sowie die 5-Sternebewertungen extrahiert und korreliert. Dafür verglichen wir die mit dem oben beschriebenen Emotionsmodell kategorisierten Emojis mit der jeweiligen 5-Sternebewertung einer Rezension. Der Spearmans Rangkorrelationskoeffizient lag bei 0.681, woraus eine starke Korrelation zwischen der Sterne-Bewertung und dem Sentiment-Anteil der Emojis folgt. Damit sind Emojis als Bewertungsindikator vergleichbar zur Sternbewertung und können somit ergänzend oder auch alternativ genutzt werden.

5.2 Beispielhafte Untersuchung von Apps auf kurzzeitige Trends

Erste Erfahrungen mit unserem Prozess haben wir bei der Untersuchung von Apps aus dem App Store gesamt-

melt. Es gibt zahlreiche Apps, bei denen plötzliche Bewertungseinbrüche existieren bzw. es spontane Anstiege in der Menge an Feedback gibt. Dies gilt insbesondere für solche Themen, die unser Ansatz erkennt.

Die App Snapchat beispielsweise hatte am 27.7.2017 eine hohe Menge an Feedback. Ursache war eine temporäre Störung des Dienstes. Einen ähnlichen Bewertungsverlauf konnten wir Ende Juni 2017 beobachten, als die App ein Update erhielt, das es ermöglichte, seine Freunde auf einer Kartenansicht darzustellen. Die Nutzer schienen nicht zu wissen, wie die Funktion zu benutzen ist und es gab einen Einbruch der Bewertung sowie verstärkt negative Emojis.

Des Weiteren untersuchten wir Instagram, bei dessen Bewertungsverlauf es zwei deutliche Einbrüche Anfang und Ende Juli 2017 gab, die gleichzeitig mit einer erhöhten Menge an Feedback und vielen negativen Emotionen einhergingen. Bei der ersten Situation handelte es sich um ein kurzzeitiges Problem des Dienstes, wo Benutzerkonten irrtümlich als gelöscht markiert wurden und so eine Nutzung des Dienstes nicht mehr möglich war. Die zweite Situation war eine etwa zwölf Stunden andauernde Downtime für einen Teil der Nutzer.

Die aufgezeigten Beispiele machen deutlich, dass Probleme nicht zwangsläufig rein in der Mobilanwendung zu verorten sind. Apps sind immer mehr Teil von größeren Systemen sind, die eine sofortige Reaktion auf Probleme erforderlich machen.

6. Fazit und Ausblick

Bei den Forschungen zu Opti4Apps haben wir die Notwendigkeit erkannt, leichtgewichtige Textanalysemethoden zu entwickeln. Dazu haben wir einen Qualitätssicherungsprozess definiert. Ein Kernbestandteil des Prozesses bildet die Analyse von Emojis innerhalb des Feedbacks und die schnelle Erkennung von Trends, um auf dieser Basis Handlungsempfehlungen ableiten zu können.

Damit dies möglich ist, haben wir ein Emotionsmodell erstellt und eine Umfrage zu Emojis mit über hundert Teilnehmern durchgeführt. Kernergebnis der Umfrage ist, dass Menschen eine sehr ähnliche Wahrnehmung von Emojis haben. Folglich können große Mengen an Texten auf Grund der verwendeten Emojis schnell analysiert werden. Dies ermöglicht eine umfangreichere Kategorisierung von Texten, verglichen mit der regulären 5-Sternebewertung, die beispielsweise im Apple App Store üblich ist.

Die Ergebnisse flossen in unseren Qualitätssicherungsprozess zur Trenderkennung in textuellem Feedback ein. Zentrale Anforderung an den Prozess sowie an

die technische Umsetzung ist eine einfache Bedienbarkeit und die Einsatzfähigkeit für hochiterative Produktentwicklung.

Die Analyse von Emojis erlaubt uns überall Feedback zu analysieren, wo keine bewerteten Metriken, wie Bewertungsterne oder Zustimmungen, mit dem Feedback mitgeliefert werden. Dies trifft insbesondere auf Feedback im Supportbereich und soziale Medien zu.

Der Vergleich von Sternebewertung und des emotionalen Sentiments zeigt eine starke Korrelation der beiden Metriken auf. Damit können wir eine stärker automatisierte Auswertung von Nutzerfeedback in heterogenen Datenquellen realisieren. Darüber hinaus haben wir in Form einer Post-Mortem Analyse Beispiele aufgezeigt, wie unser Prozess bei realen Produkten und deren Feedback gewirkt hätte. Auch wenn die Evaluationen auf Apps beschränkt war, eignet sich der Ansatz auch für andere Produkte, die schnelle Innovationszyklen haben und Feedback der Nutzer zur Verfügung steht.

Einer der nächsten Schritte ist Feedback über verschiedene Datenquellentypen wie App Stores, soziale Medien und Kundenforen hinweg gezielt zu vergleichen. Da die Menge an Emojis stetig weiterwächst, haben wir eine Folgeumfrage in Arbeit. Im Wesentlichen besteht diese aus der Klassifizierung der im Jahr 2017 neu eingeführten Emojis, sowie detaillierten Fragen zu den schwer einordbaren Emojis.

Danksagung

Dieser Beitrag wurde erstellt im Kontext des Projekts Opti4Apps, gefördert durch das Bundesministerium für Bildung und Forschung (Förderkennzeichen: 02K14A182).

Referenzen

- [1] F. Elberzhager, K. Holl und S. A. Scherr, „Mobile App Testing? Nutzerfeedback automatisiert miteinbeziehen,“ *Softwaretechnik-Trends*, Bd. 37, Nr. 1, 2017.
- [2] S. A. Scherr, F. Elbertshager und K. Holl, „An automated feedback-based approach to support mobile app development,“ in *Proceedings - 43rd Euromicro Conference on Software Engineering and Advanced Applications, SEAA 2017*, Vienna, 2017.
- [3] P. Ekman und W. V. Friesen, „Constants across cultures in the face and emotion,“ *J. Pers. Soc. Psychol.*, Bd. 17, Nr. 2, pp. 124-129, 1971.
- [4] R. S. Lazarus, „Emotion & Adaption,“ 1991.
- [5] R. Plutchik, „A General Psychoevolutionary Theory of Emotions,“ in *Theorie of Emotions*, 1980, pp. 3-33.
- [6] R. Feldman, „Techniques and applications for sentiment analysis,“ *Communications of the ACM*, Bd. 56, Nr. 4, 2013.