スマートフォンアプリケーションのレビューにおける苦情の分析
-同一アプリを対象とした場合の地域による傾向の違いに対しての考察―

2017SE036 小井土文哉　2017SE110 吉田稜
指導教員：横森励士

1　はじめに

スマートフォンアプリケーションを開発する際，ユーザーが投稿するレビューは，開発者にとってユーザーからの重要なフィードバックとみなすことができる．過去の研究において，Khalid ら [1] は，北米のスマートフォンアプリケーションを対象に低評価のユーザーレビューの中の苦情の中身を分類し，最も多く発生する苦情が「機能エラー」であることを確認した．我々の研究グループでは，日英米の 3 カ国のアプリストアでそれぞれレビューの苦情内容の調査を行ったが [3][4][5]，レビューにおける傾向の違いが選択したアプリによるものかはまだ検証していない．本研究では，日英米 3 地域で共通して配布されているアプリケーションを対象に，低・中・高評価におけるユーザーレビューの苦情の中身を分類し，分類結果に地域差が存在するかを確認する．各国のユーザーレビューにおいて苦情が出現する割合，各評価帯におけるレビューでの苦情の出現頻度，出現した苦情の傾向の違いを調査する．苦情の傾向調査アプリを一致させ，全体の傾向とアプリ固有の違いを調べることで，保守活動にどのように活用することができるかを考察する．

2　研究背景

2.1　スマートフォンアプリケーションにおけるレビュー

スマートフォンアプリはアプリケーションストアを介して配布されることが一般的である．利用者はアプリケーションについての評価を配布元のストアに投稿できる．投稿したレビューは，タイトル，星評価，アプリに対してのコメントで構成されている．それらの情報は他の利用者が参考にできる情報であるとともに，開発者へのフィードバックとみなすことができる．スマートフォンアプリを開発するにあたって，ユーザーレビューは開発者が方向性を決定する際に重要な情報になっている．

2.2　アプリケーションのレビューを対象とした分析

Leonard Hoon[2] は App Store の 17,330 のアプリから 870 万件のレビューにおいて使われている単語をすべて抽出し，評価との関連を調査した．その結果，肯定的な意味を表す単語より否定的な意味を表す単語の方が出現する種類数が多く，利用者は不満点を感情とともに独自の表現で伝える傾向があることが分かった．

Khalid[1] は欧米で提供されている無料 iOS アプリを対象に低評価レビューを分析をし，どのような種類のコメントが多くされ，どのような種類のコメントが低評価につながりやすいのかを調査した．低評価ユーザーレビューにはコメントの内容に基づき， 表 1 で示す 12 種類の苦情を表現するタグが付けられた．各苦情タイプについて低評価レビューの中でどれだけ出現しやすいか（苦情頻度）が集計され，その結果，最も多く発生する苦情が「機能エラー」,「機能要求」,「強制終了」であった．これらの情報を Khalid らは改善を目的としたリソースを配分する時に役立つ情報であると結論づけている．

安部 [3] らは日本のアプリケーションに対して Khalid と条件を揃えた上で低評価レビュー分類を行った．共通点から世界的にユーザーが考えていることを，相違点から日本のユーザー固有の特徴が得られると考え，日本向けアプリを開発する際，特別に考えなければならない事を提言できると考えた．結果は北米の利用者に対する分析と同様に，最も多い苦情は「機能エラー」に関することであった．相違点として，北米では「機能エラー」に次いで「機能要求」が多かったのに対し，日本では「魅力がない」などのソフトウェアの改善につながらない提言が多かった．

平井 [4] らは低評価のレビューだけでなく中高評価のレビューにも提言というかたちで苦情が存在すると考えた．そこで日本のアプリについて中高評価レビューも含めたユーザーレビュー全体を対象とし，苦情内容の分析を行った．結果として，中評価レビュー内には 8 割，高評価レビュー内にも 3 割のレビューに，苦情に相当する内容の提言が書かれていた．その中には「機能要求」が多く存在し，提言を得るために中高評価レビューを見る価値があることが分かった．松永 [5] は平井らの研究をふまえて，米英でも低評価だけでなく中・高評価にも提言というかたちで苦情が存在し，それらは「機能要求」などの提言であることが一般的な傾向かを調査した．過去の研究と同じ条件になるようアプリを選択し，レビューの分析を行った．その結果，低評価レビューには「機能エラー」の意見が多く存在し，中高評価レビューにも「機能要求」などの意見が上位を占め，同様な傾向が見られることが記された．米国では，高評価レビューの中に提言がより多く存在し，自らの意見がはっきりと述べられている内容が多かった．

3　同一アプリを対象とした場合の地域によるユーザーレビューの苦情の傾向の違いの調査

3.1　研究動機

過去の研究により，苦情内容の分布には同様の傾向が存在した上で，苦情内容の分布に地域差が存在することも分かった．今までの分析では地域ごとに苦情の分析対象のアプリケーションが異なるので，今まで確認できた差がアプ

表 1 Khalid らの分析における苦情の種類 ([1]p.74 の表 2 を和訳)

苦情タイプ	苦情の詳細	レビュー例
強制終了	アプリケーションが強制終了する	起動後，すぐに落ちる
互換性	特定のデバイスや OS のバージョンに問題がある	ipod touch では画面の半分しか見れない
機能削除	特定の機能がアプリケーションを台無しにしている	アプリ自体は素晴らしいが広告を除いてほしい
機能要求	より良くなるために，機能を追加する必要があると感じている	アラートを設定できる機能がない
機能エラー	アプリケーションの特定の問題に言及し，不満を感じている	アプリケーションを開かないと通知が来ない
隠されたコスト	全てを経験するために追加の隠されたコストが必要	リアルマネーを使い，コインの購入を強いてくる
インターフェース設計	デザイン，制御，映像について不満がある	デザインが小奇麗でなく，わかりづらい
ネットワーク問題	ネットワークに問題があるか，応答速度が遅い	新しいバージョンがサーバーにつながらない
プライバシーと倫理	プライバシーを侵す，または反倫理的である	あなたとの接触が目的なアプリケーション
アプリが応答しない	入力の応答が遅い，または全体的に遅い	古いバージョンに戻したい！スクロールが遅い
魅力のない内容	特定のコンテンツが魅力的ではない	画面の見栄えは良いが，退屈でつまらないゲーム
重いリソース	アプリケーションがバッテリーまたは容量を消費しすぎる	常時 GPS を使い，バッテリーが消費される
特定できない	ただ単にアプリケーションが悪いと言っている	正直なところ，最悪のアプリケーション

リケーションの違いによるものかは，検証されていない．

3.2 研究目的

本研究では，同時期に公開されているアプリケーションはほぼ同一の内容を有していると仮定し，日米英の 3 地域において公開されている同一のアプリケーション群に対してレビュー取得をし，それらの苦情内容の分類を行う．これにより，各地域のユーザーがどのような点に着目してアプリケーションを評価しているか調査する．全体の傾向と個別のアプリの違いを調べることでどのような点が全体の傾向として現れるか調査する．各国の苦情の差をどのようにアプリケーションの開発や品質の向上に役立てることができかを考える．3 地域で同一のアプリケーションを対象とした場合でも、過去の分析と同様の傾向が得られることを示すことで，今までに得られた分析の成果における傾向の一般性を補強しする．これらを調査することでスマートフォンアプリケーションの保守におけるレビューの活用方法を確立することを目的とする．

4 評価実験

4.1 対象アプリケーションの選出

日本・米国・英国の「iTunesStore」を調査し，3 地域における平均星評価が 4.0 以上の高評価アプリと，それ以下の低・中評価のアプリを 10 個ずつ選出し，3 か月間レビューの取得を行った．アプリごとに信頼水準 95 %，信頼区間 5 %でレビューの抽出件数を決定し，抽出したレビューについて苦情内容の分類を行った．

4.2 調査手順

[3]，[4]，[5] と同様に，以下の手順で調査を行った．苦情内容を分類した結果，表 1 の区分で分類出来なかった苦情は存在しなかった．

1. 各国の App store において公表されているアプリケー ションのユーザーレビューを「iTunes Store Web Service Search API」を通じて一定期間取得する．
2. 各ユーザーレビューの id，レビュー タイトル，星評価，コメントを抽出する．
3. 低，中，高評価のユーザーレビューの苦情を分類し，低，中，高評価それぞれのレビュー中の苦情の中で各種類の苦情がどれだけ存在するか（苦情中の占有率）を調査する．

4.3 分析結果の紹介

高評価・低中評価それぞれから 3 件ずつ，表 2 で示す計 6 件のアプリケーションを対象として，苦情レビューにおける各苦情項目の占有率を求めた．表 2 では，アプリケーション名と，属するジャンル，日英米での今までに投稿された総レビュー数，3 地域での平均星評価を表に示す．分析した 6 つのアプリから PINTEREST における結果を紹介する．最初に，3 カ国の苦情の出現頻度を表 3 にまとめた．どの評価帯にも苦情は存在し，評価が上がってくるにつれて出現割合は減少している．同様に 3 カ国の PINTEREST の苦情の分類結果を表 4，5，6 にまとめた．日英では，低評価で「強制終了」や「アプリが応答しない」などの意見が多く見られた．米国では，特定できたレビューのタイプ内では「機能要求」「機能エラー」が多く見られた．「機能要求」では「写真を並べ替える機能」や「保存機能などをつけてほしい」という意見が多く述べられていた．低評価に多く見られた「機能削除」では，「広告を削除してほしい」という意見が多くを占めた．

4.4 その他アプリにおける結果の要約

- 「super mario run」
  3 カ国共に低評価で「課金をしないと先のストーリーに進めない」など，アプリ内での課金にかかる「コスト」についての提言が多く見られた．日英の高評価では，「機能要求」，「隠されたコスト」が高いのに対し，米国では高評価に「機能要求」の意見が多く見られ，「十字の操作ボタンをつけてほしい」などの動作をスムーズに行うための要求が多く見られた．
- 「Twitter」
  低中評価帯では，日英共に「機能要求」，「機能エラー」が多く，高評価帯では，「機能要求」についての提言が多く見られた．米国と比べて日英では，単に「酷い」

表 2 分析対象のアプリケーション

アプリケーション	ジャンル	日レビュー	米レビュー	英レビュー	日平均	米平均	英平均
Twitter	ニュース	121 万	290 万	49.5 万	4.3	4.7	4.7
Pinterest	ライフスタイル	19 万	320 万	35.9 万	4.6	4.8	4.7
Mario kart	ゲーム:アクション	41 万	9.22 万	264 万	4.6	4.7	4.7
Netflix	エンターテインメント	1.4 万	19.6 万	2.7 万	3.0	4.1	4.0
Amazon Prime Now	フード・ドリンク	1.9 万	10.4 万	1090	2.5	3.2	3.0
Super Mario run	ゲーム・アクション	1.9 万	5800	13.1 万	3.1	3.6	4.7

表 3 レビューにおける苦情の出現割合 (PINTEREST)

	星 1	星 2	星 3	星 4	星 5
US	0.97(177/182)	0.91(122/134)	0.82(134/163)	0.70(115/164)	0.30(72/1238)
GB	0.97(1290/133)	0.93(70/75)	0.81(76/94)	0.64(72/112)	0.20(32/161)
JP	1(169/169)	0.92(94/102)	0.86(95/111)	0.76(84/110)	0.46(53/114)

表 4 PINTEREST の分類結果 (米)

US	星 1	星 2	星 3	星 4	星 5
強制終了	3(0.02)		2(0.01)	1(0.01)	2(0.03)
互換性	5(0.03)	4(0.03)	11(0.08)	8(0.07)	2(0.03)
機能削除	34(0.19)	29(0.24)	22(0.16)	19(0.17)	2(0.03)
機能要求	43(0.24)	35(0.29)	42(0.31)	46(0.3)	18(0.25)
機能エラー	28(0.26)	24(0.20)	28(0.21)	20(0.17)	8(0.11)
隠されたコスト					
インターフェース設計	11(0.06)	10(0.08)	9(0.07)	5(0.04)	
ネットワーク問題	2(0.01)	1(0.01)			
プライバシーと倫理	5(0.03)	1(0.01)	2(0.01)	3(0.03)	2(0.03)
アプリが応答しない	17(0.10)	14(0.11)	10(0.07)	7(0.06)	6(0.08)
魅力のない内容	7(0.04)	1(0.01)	3(0.02)		1(0.01)
重いリソース			1(0.007)		1(0.01)
特定できない	22(0.12)	3(0.02)	4(0.03)	6(0.05)	30(0.41)
苦情件数	177	122	134	115	72
苦情なし	5	12	29	46	166
合計	182	14	163	161	238

表 5 PINTEREST の分類結果 (英)

GB	星 1	星 2	星 3	星 4	星 5
強制終了	31(0.24)	12(0.17)	2(0.03)	5(0.07)	3(0.09)
互換性	8(0.06)	6(0.09)	1(0.01)		
機能削除	11(0.09)	12(0.17)	10(0.13)	18(0.25)	5(0.16)
機能要求	24(0.19)	12(0.17)	26(0.34)	18(0.25)	11(0.34)
機能エラー	18(0.14)	10(0.14)	10(0.13)	9(0.13)	1(0.03)
隠されたコスト	1(0.01)				
インターフェース設計	6(0.05)	4(0.06)	4(0.05)	3(0.04)	2(0.06)
ネットワーク問題	1(0.01)	2(0.03)		2(0.03)	
プライバシーと倫理				1(0.01)	
アプリが応答しない	20(0.16)	11(0.16)	19(0.25)	10(0.14)	2(0.06)
魅力のない内容	1(0.01)	1(0.01)	1(0.01)		
重いリソース	1(0.01)		2(0.03)		
特定できない	7(0..05)		1(0.01)	6(0.08)	8(0.25)
苦情件数	129	70	76	72	32
苦情なし	4	5	18	40	129
合計	133	75	94	112	161

表 6 PINTEREST の分類結果 (日)

JP	星 1	星 2	星 3	星 4	星 5
強制終了	25(0.15)	23(0.24)	30(0.32)	24(0.29)	12(0.23)
互換性	8(0.05)	9(0.10)	2(0.02)	8(0.10)	2(0.04)
機能削除	15(0.09)	8(0.09)	6(0.06)	7(0.08)	2(0.04)
機能要求	25(0.15)	10(0.11)	25(0.26)	19(0.23)	11(0.21)
機能エラー	23(0.14)	15(0.16)	9(0.09)	5(0.06)	6(0.11)
隠されたコスト					
インターフェース設計	13(0.08)	7(0.07)	5(0.04)	4(0.05)	2(0.04)
ネットワーク問題	3(0.02)				
プライバシーと倫理	7(0.04)				1(0.02)
アプリが応答しない	43(0.25)	20(0.21)	15(0.16)	16(0.19)	13(0.25)
魅力のない内容		1(0.01)			
重いリソース			2(0.02)		
特定できない	7(0.04)	1(0.01)	2(0.02)	1(0.01)	4(0.08)
苦情件数	169	94	95	84	53
苦情なし		8	16	26	61
合計	169	102	111	110	114

など具体性のない提言が多かった．米国では，低中高評価共に「機能要求」に関しての提言が多く見られ，具体的にどのような機能を求めているかが書かれていた．また日本と比べて，英米での低中評価帯で「人種差別」などの「プライバシー・倫理」が最も多く存在することが分かった．

- 「Netflix」
  3 カ国ともに「機能要求」「機能エラー」が上位を占め，「ジャンル検索機能を増やしてほしい」や「履歴を見られるようにしてほしい」といった声が多く見られた．また米国では「CUTIES」という映画のポスターや説明文を性的に描いているといった「プライバシーと倫理」に該当するレビューが多く見られた．日英では特に「プライバシーと倫理」に該当するレビューは見受けられなかった．
- 「Amazon Prime Now」
  3 カ国共に各評価帯で「機能要求」が上位を占める結果になった．主に，「利用可能エリアを広げてほしい」という機能の改善を要求するレビューが多く見られた．次いで「機能エラー」，「アプリケーションが応答しない」が多く見られるという 3 カ国共通の傾向が見られた．さらに，日英において各評価帯で見られた「隠されたコスト」という課金アプリならではの苦情レビューは，米国では低評価帯でのみ見受けられた．

- 「mario kart」
  「マップを増やしてほしい」や「車のカスタマイズを増やしてほしい」など「機能要求」「機能エラー」が 3 カ国上位を占めるとともに，「ガチャの排出率がおかしい」などといったコスト面の意見が多く見られた．米国では高評価帯で「機能要求」が多く見られた．

5 考察

5.1 3 カ国の傾向の共通点

低評価帯では，「機能エラー」に加えて「機能要求」が上位を占めており，高評価帯では変わらず「機能要求」が最も多く存在した．これらの傾向は [5] での分析結果と同じ傾向であると考えられる．固有の問題としては，アプリ内で発生するコストについての苦情が見られることなど，一部のアプリにおいて特定の項目の苦情がみられる場合もあり，その場合は 3 地域ともに同じ傾向が増加していた．

5.2 3 カ国の傾向の相違点

3 カ国とも苦情種類の割合の傾向に大きな差はないが，日英の低評価レビューでは，「酷い」「ごみ」などのどの部分の改善にも繋がらない意見が多く見られる一方，米国では，[5] の結果と同様に高評価帯で多くの提言として捉えられる意見が見られた．米国では，高評価帯でも「機能要求」の割合が多くを占めていることにより，アプリケーションの品質向上に積極的であるという傾向が見られた．米英では，黒人差別などのコメントが見られ，地域向けでのコンテンツで倫理的な問題の苦情が見られた．このような地域固有の問題は地域差として現れると考えられる．

5.3 ソフトウェアの保守

見る評価帯によって苦情の傾向が違うことから，低評価帯は問題が起こっているかどうかの観測を目的として利用することが望ましいこと，高評価帯はアプリケーションの改善の方向性を決める指標として有効であることが分かった．また，同一アプリにすることにより苦情内容が地域差かアプリ固有のものか明確にすることができ，地域差は自分のアプリを海外に向けて展開しようとしたときに意識する必要があり，その指標になると考えられる．

5.4 ソフトウェア品質モデルの分類

JIS X 25010 ではソフトウェアの品質モデルを，「製品品質モデル」と**「利用時の品質モデル」**の 2 つの品質モデルに定義している．「製品品質モデル」は製品品質の特性を「機能的合性」「性能効率性」「互換性」「使用性」「信頼性」「セキュリティ」「保守性」「移植性」の 8 つに分類しており，**「利用時の品質モデル」**を**「有効性」「効率性」「満足性」「リスク回避性」「利用状況網羅性」**の 5 つに分類している．品質モデルと Khalid の苦情の種類がどの特性に当てはまるかを考察した結果が表 7 である．いくつかの項目が 1 つの苦情タイプに含まれおり，内容を精査することで，より細分化する余地があると考えられる．

表 7 表 1 の苦情の種類と製品品質モデルと**利用時の品質モデルの関係**

苦情タイプ	
強制終了	機能適合性・**有効性・満足性** **利用状況完全性**
互換性	互換性・**満足性・利用状況完全性**
機能削除	機能適合性・使用性 ・**満足性**
機能要求	機能適合性・機能効率性・使用性・**満足性**
機能エラー	機能適合性・信頼性・**有効性・満足性** **利用状況完全性**
隠されたコスト	性能効率性・**有効性・効率性**
インターフェース設計	使用性・**有効性・満足性**
ネットワーク問題	機能適合性・信頼性・**有効性・満足性**
プライバシーと倫理	セキュリティ ・**満足性・リスク回避性**
アプリが応答しない	機能適合性・信頼性・**有効性・満足性** **利用状況完全性**
魅力のない内容	性能効率性・使用性・**満足性**
重いリソース	性能効率性・**効率性**
特定できない	

6 まとめ

本研究では日英米のアプリケーションのユーザーレビューを調査し比較を行った．全体としての傾向は過去の研究と同じような結果となり，分類によりこれまでに得られた性質は，より一般的な性質であると考えられる．ただし地域固有の問題やアプリ固有の問題など，中央値の比較では出てこないような性質も見られることが分かった．今後の課題として品質モデルに基づいて分類を行うことで，さらなる知見が得られるかを確認したい．

参考文献

[1] Hammad Khalid, Emad Shihab, Meiyappan Nagappan, Ahmed E. Hassan：“What Do Mobile App Users Complain About?”, In IEEE Software, Vol.32, No.3，pp.70-77，2015.

[2] Leonard Hoon，Rajesh Vasa，Jean-Guy Schneider，Kon Mouzakis：“A Preliminary Analysis of Vocabulary in Mobile App User Reviews”，Swinburne University of Technology Faculty of Information and-Communication Technologies，pp.245-248，2012.

[3] 安部寛生，波多野雅信，小林佑汰：“日本のスマートフォンアプリケーションにおける評価の低いユーザーレビューでの苦情内容の分析”, 南山大学理工学部 2017 年度卒業論文

[4] 平井賢人，稲垣絢也：“スマートフォンアプリケーションにおけるユーザーレビューの内容の分析ー低評価レビューと高評価レビューの傾向の違いについてー”，南山大学理工学部 2018 年度卒業論文

[5] 松永夏季”スマートフォンアプリケーションのレビューにおける苦情の分析—地域による傾向の違いの調査—”南山大学理工学部 2019 年度卒業論文