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
―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―

2015SE014 平井賢人　2015SE021 稲垣絢也
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関係者にとって，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．Khalid ら [2] や，安部ら [3] はスマートフォンアプリケーションを対象に，低評価 (星 1〜2) のユーザーレビューを分析し，レビューにおいてどのような苦情が多く報告されるか，低評価をつけられやすい苦情は何かを調査した．これらの分析では低評価レビューのみを分析対象としていたが，高評価のレビューにおいても要望が提言としてこれらの意見が存在すると考えられる．本研究では，ユーザーレビュー全体を対象に同様の分析を行い，[2] や [3] の結果と比較し，ユーザーレビューにおいて苦情が出現する割合，低評価レビューと全評価レビューでの苦情の傾向の比較，各評価レビューの苦情の出現頻度を調査する．それらの情報は保守活動においてどのように活用することが妥当かを考察し，保守活動に役立てる方法について提案する．

2　背景技術

2.1　スマートフォンアプリ開発とユーザーレビュー

スマートフォンアプリケーションの開発では，短期間で開発の方向性を決定し，リリースを迅速に行うことが求められており，ユーザーからのフィードバックは，方向性を決定する上で有力な判断材料の一つと考えられる．ユーザーは使用したアプリケーションについての評価をアプリケーションストアに投稿することができる．多くのレビューは星評価と具体的なコメントで構成され，開発者は利用者がアプリケーションに対して評価している点と不満に感じている点の両方の情報を得ることができる．それらの情報は，開発者へのフィードバックや他のユーザーへのアドバイスとなっている．このように，ユーザーレビューはアプリケーションを開発する指針として重要になる．

2.2　ユーザーレビューを分析した研究

Leonard Hoon ら [1] 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．否定的な意味を表現するために使用される言葉のレパートリーは肯定的な感情が表現される場合のそれよりも有意に高いなどの調査結果が得られている．一方で，アプリケーションの苦情レビューの中身を精査することで，ユーザーがどのような要素に対して，不満をもちやすいかを分析した Khalid らによる研究がある [2]．[2] では，北米で提供されている無料 iOS アプリケーションを対象として，低評価ユーザーレビューの中でどのようなコメントが多くなされていたか，どのような種類のコメントが低評価につながりやすいかを調査した．低評価ユーザーレビューにはコメントの内容に基づいて，表 1 で示す 12 種類の苦情を表現するタグが付けられ，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）を求めている．表 2 はその結果で，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．

安部ら [3] は，日本のアプリケーションに対して [2] と条件をそろえた上で低評価レビューを分類を行い，共通点から世界的にユーザーが共通して考えていることと，相違点から日本のユーザー固有の特徴が得られるのではないかと考え，日本向けにアプリケーションを開発する場合に特別に考えなくてはならないことを提言できると考えた．表 2 はその結果で,「機能エラー」などの欠陥に関する苦情の割合は変わらない一方,「機能要求」などの提言は少なく,「魅力のない内容」など見切りをつける苦情が多かった．

3　本研究について

3.1　過去の問題点

[2]，[3] の評価では，低評価ユーザーレビューのみを分類していたが，高評価のユーザーレビューにおいても“こうすることでもっと良くなる”という形で，提言のような不満が述べられていることは少なくない．これらは単なる不満より活用できると考えられ，分類する価値があると考えた．本研究では，低評価ユーザーレビューだけでなく，ユーザーレビュー全体を分析対象として分類を行い，どのように結果が変化するかを調査する．ユーザーレビューをどうソフトウェア開発に役立てるべきかを考察する．

3.2　調査項目

高評価ユーザーレビューを含めたユーザーレビュー全体を対象として苦情内容の分類を行い，以下を調査する．

1. 評価ごと (高評価，中評価，低評価) にユーザーレビューを分け，苦情が存在する割合を求める．これにより，中評価，高評価のユーザーレビューにも一定以上要望のかたちで不満が存在することを示す．
2. ユーザーレビュー全体から苦情を抽出した場合と低

表 1 Khalid らの分析における苦情の種類 [2]

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	アプリケーションが特定のデバイスや OS のバージョンに問題がある	私の ipod touch ではアプリケーションの半分しか見れない
機能削除	1つあるいは多くの特定の機能がアプリケーションを台無しにしている	このアプリケーションは素晴らしいが広告を取り除いてほしい
機能要求	アプリケーションがより良い評価を得るために，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	アプリケーションの全てを経験するために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	アプリケーションの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	アプリケーションが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	アプリケーションが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	アプリケーションの入力の応答が遅い，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

表 2 低評価レビューにおける苦情の分類結果

	苦情頻度			
	北米のアプリ [2]		日本のアプリ [3]	
苦情タイプ	順位	中央値（%）	順位	中央値（%）
機能エラー	1	26.68	1	31.71
機能要求	2	15.13	7	4.74
強制終了	3	10.51	5	5.57
ネットワーク問題	4	7.39	9	0.95
インターフェース設計	5	3.44	4	7.19
機能削除	6	2.73	6	5.26
隠されたコスト	7	1.54	12	0
互換性	8	1.39	10	0.79
プライバシーと倫理	9	1.19	8	1.26
アプリが応答しない	10	0.73	2	11.45
魅力のない内容	11	0.29	3	7.75
重いリソース	12	0.28	11	0.59
特定できない	-	13.28	-	5.6

評価のユーザーレビューのみから苦情を抽出した場合の苦情の出現頻度はどう異なるのかを調査する．苦情の出現頻度がどのように変化するのかを理解することで，評価が一定以上のユーザーレビューにおける苦情とはどのような種類の苦情かを調査する．

3. 低，中，高評価のユーザーレビューをそれぞれを分類し，レビューのなかでその種類の苦情がどの程度存在するか（レビューにおける出現頻度）と，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．これにより，各評価のユーザーレビューだけを見て意見を抽出した場合，どういった苦情が把握でき，どのような苦情を見落とすことになるのかを調べる．

4 評価実験

4.1 調査対象のアプリケーション

安部ら [3] は，[2] の結果と比較を行うために [2] で選択されたアプリケーションと似た条件でアプリケーションを選び，表 3 に示すように，アプリケーション数，ジャンル数，サンプリング範囲や低評価，高評価のアプリケーション数などがほぼ一致するようにアプリケーションを選択した．本研究では，[3] の過程で選択したアプリケーションについて入手したユーザーレビューを用いる．各評価ごとに信頼水準 95 %，信頼区間 5 %で各アプリケーション毎に抽出件数を決定し，ユーザーレビューを抽出する．それぞれのユーザーレビュー毎に，表 1 に基づいてどのような種類の苦情が存在するかをタグ付けする．アプリケーション毎にタグ付けした結果を項目毎に出現割合（苦情頻度），苦情中の占有率を求める．

表 3 抽出対象となるアプリケーション

	アプリケーション	ジャンル	サンプリング	低評価（星 3.5 未満）アプリケーション数	高評価（星 3.5 以上）アプリケーション数
北米のアプリ [2]	20 個	15	264~383	10 個	10 個
日本のアプリ [3] と今回	20 個	13	255~375	10 個	10 個

4.2 調査結果

1. 苦情が存在する割合

図 1 レビューにおける苦情の出現割合

	星1,2	星3	星4,5	計
SINoALICE ―シノアリス―	1 (71/71)	0.40 (24/60)	0.17 (40/241)	0.36 (135/372)
アビスリウム タップで育つ水族館	1 (54/54)	0.89 (39/44)	0.09 (24/259)	0.33 (117/357)
楽天市場 ショッピング・通販アプリ お買い物でポイントが貯まる・使える	1 (143/143)	0.56 (18/32)	0.10 (20/194)	0.49 (181/369)
Amazon ショッピングアプリ	1 (210/210)	1 (27/27)	0.98 (42/43)	0.99 (217/218)
フリマアプリ-メルカリ フリマで かんたんショッピング	1 (136/136)	0.85 (23/27)	0.08 (16/198)	0.48 (175/361)
Google マップ - ナビ、乗換案内	1 (149/149)	0.9 (36/40)	0.25 (30/119)	0.70 (215/308)
Simeji 日本語文字入力＆きせかえ・ 顔文字キーボード	1 (142/142)	0.94 (45/48)	0.33 (48/145)	0.70 (235/335)
Clipbox	1 (125/125)	1 (42/42)	0.51 (81/159)	0.76 (248/326)
Instagram	1 (204/204)	0.88 (49/56)	0.28 (31/109)	0.85 (284/369)
スノー SNOW - 自撮り、 顔認識スタンプ、ウケるカメラ	1 (184/184)	1 (49/49)	0.58 (45/77)	0.90 (278/310)
YouTube - 公式アプリで動画と音楽	1 (274/274)	0.94 (32/34)	0.44 (24/55)	0.91 (330/363)
LINE	1 (298/298)	0.91 (31/34)	0.35 (15/43)	0.92 (344/375)
Facebook	1 (238/238)	0.63 (15/24)	0.15 (10/66)	0.80 (263/328)
Twitter	1 (315/315)	1 (25/25)	0.64 (18/28)	0.97 (358/368)
マクドナルドMcDonald's Japan	1 (231/231)	1 (14/14)	0.22 (10/46)	0.88 (255/291)
LINE MUSIC 音楽聴き放題、 シェアし放題（ラインミュージック）	1(205/205)	0.97(38/39)	0.41(45/109)	0.86(281/333)
ジーユー	1 (301/301)	0.71 (12/17)	0.31 (4/12)	0.96 (317/330)
ピッコマ 人気マンガが待てば全話 読める！毎日更新！	1 (61/61)	0.77 (33/43)	0.20 (48/245)	0.41 (142/349)
Gmail安全ですばやく整理しやすい GoogleのEメール	1 (171/171)	0.85(35/41)	0.15 (15/100)	0.71 (221/312)
niconico	1 (331/331)	0.94 (16/17)	0.58 (11/19)	0.98 (358/367)
平均	1	0.86	0.34	0.75

図 1 は 20 個のアプリケーション毎の各評価での苦情が出現する割合を示した図である．低評価はどのアプリケーションでも全てのユーザーレビューから苦情が出現し，中評価でも約 8 割の苦情が出現している一方で，高評価では約 3 割のユーザーレビューに苦情が出現することがわかる．高評価のユーザーレビューも十分苦情内容を調査する価値があることがわかった．

2. 日本のレビュー全体と低評価レビュー，北米の低評価レビューそれぞれの全体の苦情頻度との比較

20 個のアプリケーションに対し苦情頻度を苦情タイプ別に示し，[3] と比較したのが表 4 である．表では，苦情タイプ毎に 20 個のアプリケーション毎の苦情の出現割合の中央値を求め，多かったものから順番に並べている．

表 4 各苦情タイプの苦情の出現頻度 (日本のアプリ)

	星 1〜5		星 1，2	
苦情タイプ	順位	中央値（%）	順位	中央値
機能エラー	1	24.92	1	31.71
機能要求	2	11.19	7	4.74
強制終了	3	9.6	5	5.57
互換性	4	4.74	10	0.79
インターフェース設計	5	3.44	4	7.19
機能削除	6	2.8	6	5.26
アプリが応答しない	7	1.87	2	11.45
ネットワーク問題	8	0.94	9	0.95
重いリソース	9	0.91	11	0.59
魅力のない内容	10	0.56	3	7.75
プライバシーと倫理	11	0.49	8	1.26
隠されたコスト	12	0.14	12	0
特定できない	-	3.34	-	5.6

順位が上昇した苦情タイプは，低評価ユーザーレビューのみで調査した場合よりも，出現頻度が高い．特に「機能要求」,「互換性」に関しては 3 以上順位が上昇しており，高評価ユーザーレビュー内で高い頻度で出現することがわかった．反対に「アプリが応答しない」,「魅力のない内容」,「プライバシーと倫理」は中評価以上のユーザーレビューでは出現頻度が低いことが確認できた．ユーザーレビュー全体を調査した場合，低評価 (星 1，2) と比べ多くの項目では順位は大きく変動しなかった．これらの苦情タイプは低評価，高評価に関わらず偏りなく各評価のユーザーレビューで近い苦情頻度だということが考えられる．表 5 は [2] と本研究の結果を比較している．本研究のユーザーレビュー全体を調査した場合，最も多く報告された上位 3 つの苦情タイプは，「機能エラー」,「機能要求」,「強制終了」であり，[2] の上位 3 位までの順位と同じ結果となった．[3] の結果よりも [2] の結果に近いことがわかる．そのため，日本のユーザーレビューから北米の低評価ユーザーレビューと同等の提言を得るためには，低評価だけではなく全体を調査する必要があることがわかった．

3. 各評価における苦情の出現頻度の違いについて

表 6 は各苦情数を全苦情数で割った値の中央値を苦情タイプ別に示し，各評価を比較した表である．この表は各評価それぞれでどの苦情タイプが出現しやすいかを求めることができる．低評価では「アプリが応答しない」,「魅力のない内容」に該当するユーザーレビューが多く，中高評価では特に「機能要求」に該当するユーザーレビューが多数出現した．各評価ごとで比較した結果,「アプリが応答しない」,「魅力のない内容」,「機能要求」以外の多くの項目では順位は大きく変動しなかった．これらの苦情タイプは低評価，高評価に関わらず偏りなく各評価のユーザーレビューで近い苦情頻度だということがわかる．「機能要求」,「互換性」に関しては 3 以上順位が上昇しており，高評価ユーザーレビュー内で高い頻度で出現することがわかった．反対に「アプリが応答しない」,「魅力のない内容」,「プライバシーと倫理」は中評価以上のユーザーレビューでは出現頻度が低いことが確認できた．また「機能要求」,「互換性」,「重いリソース」は中，高評価での中央値が最も高いため，これらの苦情の提言を得るには中，高評価に着目すべきである．

表 7 は各苦情数を全ユーザーレビュー数で割った値の中央値を苦情タイプ別に示し，各評価を比較した表である．苦情タイプ毎に出現頻度が多い評価の中央値を太字にした．この表は各苦情タイプがどの評価で出現しやすいかを

表 5 各苦情タイプの苦情の出現頻度 (日本と北米のアプリ)

	日本のアプリ (星 1〜5)		北米のアプリ (星 1，2)	
苦情タイプ	順位	中央値（%）	順位	中央値
機能エラー	1	24.92	1	26.68
機能要求	2	11.19	2	15.13
強制終了	3	9.6	3	10.51
互換性	4	4.74	8	1.39
インターフェース設計	5	3.44	5	3.44
機能削除	6	2.8	6	2.73
アプリが応答しない	7	1.87	10	0.73
ネットワーク問題	8	0.94	4	7.39
重いリソース	9	0.91	12	0.28
魅力のない内容	10	0.56	11	0.29
プライバシーと倫理	11	0.49	9	1.19
隠されたコスト	12	0.14	7	1.54
特定できない	-	3.34	-	13.28

表 6 各評価における苦情タイプ毎の占有率

	星 1，2		星 3		星 4，5	
苦情タイプ	順位	中央値（%）	順位	中央値	順位	中央値
機能エラー	1	31.71	2	23.26	2	15.25
アプリが応答しない	2	11.45	8	1.44	9	0.70
魅力のない内容	3	7.75	10	0.70	11	0
インターフェース設計	4	7.19	3	6.98	3	6.05
機能削除	5	5.57	5	4.36	4	4.76
強制終了	6	5.26	4	6.28	5	4.46
機能要求	7	4.74	1	24.89	1	37.58
プライバシーと倫理	8	1.26	11	0.60	10	0.39
ネットワーク問題	9	0.95	7	2.84	7	1.27
互換性	10	0.79	6	4.01	6	3.39
重いリソース	11	0.59	9	1.17	8	0.76
隠されたコスト	12	0	12	0	11	0
特定できない	-	5.60	-	1.79	-	3.35

表 7 各評価におけるユーザーレビューの苦情頻度

	星 1，2		星 3		星 4，5	
苦情タイプ	順位	中央値（%）	順位	中央値	順位	中央値
機能エラー	1	**31.71**	2	19.60	2	4.02
アプリが応答しない	2	**11.45**	8	1.31	9	0.17
魅力のない内容	3	**7.75**	10	0.67	11	0
インターフェース設計	4	**7.19**	3	**7.18**	3	2.37
機能削除	5	**5.57**	5	4.15	4	1.64
強制終了	6	**5.26**	4	**5.93**	5	1.61
機能要求	7	4.74	1	**22.73**	1	**11.26**
プライバシーと倫理	8	**1.26**	11	0.23	10	0.13
ネットワーク問題	9	0.95	7	**2.84**	7	0.48
互換性	10	0.79	6	**4.09**	6	0.91
重いリソース	11	0.59	9	**0.71**	8	0.33
隠されたコスト	12	0	12	0	11	0
特定できない	-	5.60	-	1.01	-	1.18

求めることができる．「機能要求」は中，高評価共に 1 位で出現頻度が高かった．しかし，高評価には苦情件数自体が少ないため，中評価よりも出現頻度が低いことがわかった．その他の中高評価で高かった項目は，中評価での出現頻度が最も高く，高評価のみの確認だと苦情を見落としてしまう可能性があることがわかる．

5 考察

- 高評価には建設的な意見が多く含まれている

日本の低評価レビューからは「機能エラー」,「アプリが応答しない」,「魅力のない内容」の苦情タイプのレビューが多く出現する．しかし, それ以外は“つまらない”などの具体性のないユーザーレビューが多く出現した．一方で高評価のユーザーレビューを調査した結果,「機能要求」に関するレビューが多く出現し，今後のアプリケーションがより良くなるアイデアなど建設的な意見のレビューが多くみられた．高評価レビューの中には苦情の絶対数が少ないため，高評価に多く現れる苦情が同様に出現する，中評価以上のレビューも合わせて確認するべきである．

- 海外と日本の高評価アプリケーションの比較

海外の低評価レビューと日本の全体レビューを比較した場合，苦情の分布結果はほとんど同じとなった．一方，北米のアプリケーションを一つ選択し高評価レビューを確認したところ海外のレビューでは高評価での苦情が特に少なかった．海外の高評価はアプリケーションが問題なく動いていることに対しての高評価レビューであり，日本は高評価であっても操作性などについての不満や改善点を高評価にレビューしていることがわかった．このことから同じ調査でもレビューの意味が異なり，レビューや評価のしかたに国ごとに差があることがわかった．

- 保守作業の観点からどうレビューを利用するべきか

図 2 は各苦情タイプがどのような保守作業に関連しているのかを表した図である．修正保守であるネットワーク障害などの保守の活動は低評価によく現れ，適応保守や改善保守に該当する活動は中，高評価に現れる苦情タイプから改善につながる意見を取得できる．作業の目的に応じて調査

図 2 保守作業と苦情タイプとの関連性

保守	保守の活動	苦情タイプ	多く出現する評価
修正保守	要求仕様の変更	機能エラー　アプリ応答 強制終了　ネットワーク問題	低
修正保守	構築上の設計仕様の変更	機能エラー　アプリ応答 強制終了　ネットワーク問題	低
修正保守	プログラムの設計仕様の変更	機能エラー　アプリ応答 強制終了　ネットワーク問題	低
修正保守	プログラムコードの変更	機能エラー　アプリ応答 強制終了　ネットワーク問題	低
修正保守	ネットワークの障害対応	ネットワーク問題	低
修正保守	スクリプトのテストの修正	機能エラー　アプリ応答 強制終了　ネットワーク問題	低
適応保守	OSのアップデート	互換性	中 高
適応保守	必要のない機能の除去	機能削除	全
適応保守	ファンクションキー定義の変更	互換性	中 高
改善保守	プログラムの設計仕様の追加	機能要求	高
改善保守	テストの網羅度を向上させる テストの組み合わせ変更	機能要求， インターフェース設計 重いリソース	高
改善保守	読みやすさを強化するコードや 設計の変更	インターフェース設計 重いリソース	高
予防保守	型チェックの付加	該当なし	
予防保守	障害処理の増強	該当なし	
アプリケーション システム	利用者に魅力を感じさせる コンテンツの追加または変更	魅力のない内容	低
アプリケーション システム	倫理感に基づくコンテンツの変更	プライバシーと倫理	高

するレビューを変えることで，より効率的に保守の活動に活かせる意見を集めることができると考えられる．

6 まとめと今後の課題

本研究では，日本のアプリケーションのユーザーレビューを調査し，ユーザーレビューの評価ごとに苦情の出現頻度の割合が違うことを確認した．それらは国ごとに異なると考えられ，ユーザーレビューを活用するときに考慮が必要である．今後の課題として，他ジャンルのアプリケーションについての特徴や，より多くの国で調査し，得られた傾向の違いを紹介することで開発者がアプリケーションを開発，運用する際に参考になる情報を提供したい．

参考文献

[1] 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.

[2] 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.

[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，2018.