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
―レビューの自動分類に関する考察―

2019SE013 伊藤陽　2019SE021 紀本光琉
指導教員：横森励士　井上克郎

1　はじめに

スマートフォンアプリケーション（以下，アプリ）におけるユーザーレビューは，使用者の意見が数多く投稿され，それと同時に多くの評価情報を見ることができる．それらの評価情報は，アプリの開発者にとって今後の開発や保守における重要な参考意見となる．Khalid ら [1] や安部ら [2] はアプリを対象としてレビューの分類を行い，どのような内容の苦情が多くみられるのか，低評価に結び付きやすい苦情は何かを調査した．[1]，[2] では有益な結果が得られたが，その際分類に用いたレビューのタグ付けは手動であった．本研究では，タグ付けの作業を機械学習によって自動化できるかを検証することを目的として，先行研究 [2] で分類されたレビューを活用しながら，実際にレビューを機械学習により分類する．分類した結果から現状の課題を洗い出し，十分な精度を持った分類手法を提案することを目指す．あらかじめ手動でタグ付けを行うなどして，タグ付け済みのレビューを取得する．そのタグ付けされたレビューを用いて学習モデルを作成し，分類器を作成し機械学習を行う．十分な精度を持った学習モデルを作成し，タグ付けを自動化することで，大量のレビューを分析対象に加えることができ，より実用的な性能が得られることが期待される．

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について

アプリは通常アプリケーションストア（以下，アプリストア）を介して配布される．ユーザーは使用したアプリの評価をアプリストアに投稿できる．投稿されたレビューはタイトル，星評価，アプリに対するコメントから構成される．それらのレビューは他のユーザーが参考にできる情報であるとともに，アプリの開発者へのフィードバックとなっている．アプリの開発において，ユーザーレビューは保守や運用の方向性を決める際に重要な情報である．

2.2　アプリケーションのレビューを対象とする分析

アプリのレビューを精査することで，ユーザーがどのような要素に対して，不満を持ちやすいかを調査した Khalid らによる研究 [1] がある．[1] では，北米で提供されている無料 iOS アプリを対象に，低評価レビューでどのようなコメントが多く寄せられているのか，どのような種類のコメントが低評価に繋がりやすいのかを調査するために，レビューを苦情の種類ごとに表 2 に示す 12 種類のカテゴリーに分類し，それぞれのレビューに対して各カテゴリーを示すタグを１つ付けた．各カテゴリーについての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），ユーザーから嫌悪され，星評価１がつきやすい項目はどれか（苦情影響力）を求めている．調査結果では，苦情頻度の高い苦情として「機能エラー」，「機能要求」，「強制終了」が挙げられ，星評価１がつきやすい項目として「プライバシーと倫理」，「隠されたコスト」が挙げられた．これらの情報は改善を目的としたリソースの配分時に役立つ情報であると結論づけている．

安部ら [2] は，日本のアプリに対して Khalid らと条件を揃えたうえで，低評価レビューを分類し，共通点から世界的にユーザーが共通して考えられていることと，相違点から日本のユーザーにのみ見られる特徴が得られるのではないかと考え，日本向けにアプリを提供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．結果として，「機能エラー」などのアプリの欠陥に対する苦情の割合は変わらなかったが，「機能要求」などの苦情が少なく，低評価のユーザーが消極的であることがわかった．

平井ら [3] は低評価レビューだけではなく中・高評価のレビューにも提言という形で苦情が存在するのではないかと考えた．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，中評価レビューの中には 8 割，高評価レビューの中には 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．その中では「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わかった．

2.3　先行研究でのレビューの分類プロセス

先行研究 [2]，[3] でのレビュー取得の手順を示す．

1. 「iTunes Store Web Service Search API」を通じてアプリのユーザーレビューを取得する．
2. 各ユーザーレビューの更新日，id，タイトル，星評価，コメントを抽出する．
3. 信頼水準 95%，信頼区間 5% で各アプリでタグ付けを行うレビューの件数を決定する．この件数に基づいて，レビューを無作為に抽出する．
4. [1] で特定したカテゴリーのうち，各レビューがどのカテゴリーに含まれるかを手動で判定しタグ付けする．

表 1 Khalid ら [1] の分析におけるカテゴリーの種類

カテゴリー	カテゴリーの詳細	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	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	デザインが小綺麗ではなく，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繋がらない
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
重いリソース	バッテリーまたは容量を消費しすぎている	バッテリー消費が凄い
魅力のない内容	特定のコンテンツが魅力的ではない	退屈でつまらないゲーム
アプリが応答しない	起動しない，無反応である	押しても反応しません

3 レビューの自動分類に関する考察

3.1 研究動機

先行研究 [2]，[3] では，複数の人間がそれぞれタグ付けを行った後で，その結果を比較しながら決定するという形でレビューへのタグ付けが行われた．ただし，この方法ではタグ付け可能なレビューの件数に制約ができ，タグ付けを行う人間によってタグ付け結果に違いが生じてしまう可能性が考えられる．実用化を考えると，自動化が不可欠であると考えられ，タグ付けの自動化が可能かどうかを検討する必要がある．レビューのタグ付けを自動化することにより，大量のレビューを分析対象に加えることができる．それに伴って，機械学習における分類の精度や分類結果の評価の精度が向上することが期待できる．

そこで，本研究では，先行研究のタグ付け結果などを利用しながらレビューを機械学習により分類し，学習モデルを作成する．タグ付けの作業を機械学習によって自動化した際の学習モデルの精度を検証するとともに，どのカテゴリーの予測が正解しやすいか，間違えやすいかなどの観点から考察する．

3.2 タグ付けの分類プロセス

機械学習を用いてタグ付けを行う場合，学習モデルを作成するために一部のレビューはタグ付けが行われている必要がある．また，タグ付けは自動で行われることを考慮すると，タグ付けの件数を制限する必要はないと考えられる．そこで，手順は以下の通りとなると考えられる．

1.「iTunes Store Web Service Search API」を通じてアプリのユーザーレビューを取得する．
2. 各ユーザーレビューの更新日，id，タイトル，星評価，コメントを抽出する．
3. 学習モデルを構築するためにレビューの一部を取得し，手動でタグ付けを行う．または，タグ付けが済んでいるレビューを取得する．
4. タグ付けされたレビューを学習用データとして機械学習を行い，学習モデルを作成する．
5. タグ付けがされていないレビューについて，学習モデルに基づきタグ付けを行う．

3.3 リサーチクエスチョン

学習モデルを作成し，以下のことを調査する．

- 2000 件のレビューに対して機械学習を行うことで，結果としてどのくらいの精度が得られるのか？
- 予測が正確にできているジャンルはどのようなジャンルであるか？

3.4 モデルに用いるデータセット

安部ら [2] によってタグ付けされたレビューを用いる．単一の項目についてタグ付けが行われているレビューについて，コメント，カテゴリーを抽出し，学習データの候補となるデータを 2000 件作成した．それぞれのアプリの件数は表 2 のようになった．

表 2 使用したアプリと件数

アプリ名	件数	アプリ名	件数
LINE	66	YouTube	29
シノアリス	12	バジリスク	120
SNOW	41	FF	82
Twitter	108	Instagram	98
niconico	67	楽天市場	48
メルカリ	43	アビスリウム	57
ピッコマ	29	Facebook	85
Simeji	95	LINE MUSIC	72
Clipbox	104	ジーユー	148
Amazon	125	GoogleMap	97
マクドナルド	115	Gmail	102
アサシン	41	クロノトリガー	17
しょぼんのアクション	29	ズーキーパー DX	21
ドラゴンクエスト V	6	ミステリージャーニー	7
一筆書きパズル	14	テトリス	122
		合計	2000

3.5 機械学習に用いたアプローチについて

本研究では，Bag of Words とニューラルネットワークを組み合わせた手法 [4] を適用した．まず，学習データのコメントを one-hot ベクトル化する．さらに，ニューラルネットワークに対し，このベクトルを入力，カテゴリーを出力として学習を行う分類器を作成した．自然言語処理で使用されることが多いアルゴリズムを使って機械学習を行

う．実験ではユニット数を 100，中間層の層数を 1 層としている．

3.6 評価方法

リサーチクエスチョンを調査するために以下の 2 つの実験を行う．

- 実験 1 ：総データ数を 2000 件として，データ数を増やしながら k-分割交差検証 [5] を行う．
- 実験 2 ：2000 件を 500 件ずつに 4 等分して，1500 件の学習モデルで 500 件の予測を行う操作を 4 回行い，再現率と適合率を確認する．

4 評価実験

4.1 実験 1

データ数を 2000 件としたデータセットを用意した．このデータセットからランダムに 100 件データを選び 10 分割交差検証を行った．順次ランダムに選んだ 100 件のデータを追加し 10 分割交差検証を行った．その結果を表 3 及び図 1 に示す．学習データが 2000 件の際の精度は 0.68314 となった．ここでは精度として，判断結果がタグ付けした結果と等しかった割合を示す．実験 1 の結果からレビューの件数を増やせば精度が向上していくことがわかった．しかし，学習データが 2000 件付近では，精度の向上の割合が初期に比べて減少しており，精度の向上を期待するにはさらに多くのデータ収集の必要があると考察した．

表 3 データを 100 件ずつ増やしながら交差検証した結果

データの件数	精度 (スコア)
100 件	0.406
200 件	0.48749
300 件	0.56096
400 件	0.5745
500 件	0.589
600 件	0.60588
700 件	0.61667
800 件	0.62173
900 件	0.62661
1000 件	0.65126
1100 件	0.65723
1200 件	0.66413
1300 件	0.6673
1400 件	0.66168
1500 件	0.66407
1600 件	0.67302
1700 件	0.67677
1800 件	0.67673
1900 件	0.68759
2000 件	0.68314

4.2 実験 2

2000 件のデータを 500 件ずつ 4 等分して，1500 件の学習モデルで 500 件の予測を行う操作を 4 回行った．この 4 回の操作で 2000 件のデータすべてが 1 回ずつ分類の対象となる．表 4 は機械学習により分類した結果で，各行では，各カテゴリーにタグ付けされたデータが機械学習による分類ではどのカテゴリーに分類されたかを示している．

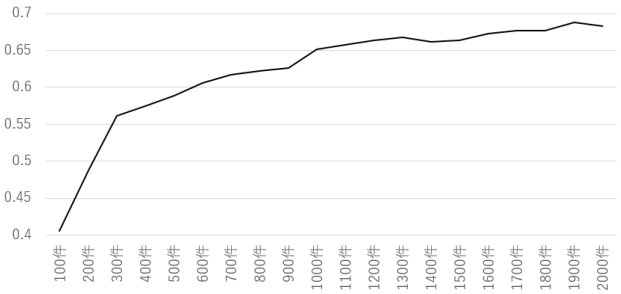

図 1 表 3 の精度の変化を表すグラフ

例えば，「強制終了」とタグ付けされた 238 件のデータのうち，194 件は正しく「強制終了」と判定されたが，18 件は「互換性」と判断されるなど，44 件のデータは正しく分類されなかったことを示している．この結果から「強制終了」とタグ付けされた 238 件のデータにおける再現率は 0.81513 であることがわかる．また，「強制終了」と予測された 231 件のデータのうち，194 件のデータのカテゴリーが「強制終了」であったが，37 件のデータが実際にはその他のカテゴリーにタグ付けされていたデータであった．この結果から「強制終了」と予測された 231 件のデータにおける適合率は 0.83983 であることがわかる．

この表から，高い再現率が得られたカテゴリーとして「強制終了」，「互換性」，「機能エラー」が挙げられる．それらはいずれもデータセット中に多く存在していることがわかる．低い再現率となったカテゴリーは，「重いリソース」，「隠されたコスト」，「プライバシーと倫理」となり，データセット中の件数も少ないものであった．データセット中の件数の多寡が再現率に影響を及ぼすと考えられる．

一方で，高い適合率が得られたカテゴリーは「プライバシーと倫理」，「重いリソース」，「強制終了」で，低いカテゴリーは，「魅力のない内容」，「隠されたコスト」，「アプリが応答しない」となった．この結果から適合率の高い低いは，そのカテゴリーであると判断可能な単語が多いかどうかに影響されると考えた．

5 考察

2000 件のデータセットに対して分析を行った結果，約 0.68 という精度が得られた．また，データ数が多いカテゴリーは正解されやすく，データ数が少ないカテゴリーは間違えられやすいという傾向が得られた．この結果を踏まえて分類精度の向上を考えた場合，現状では次のような課題が存在すると考えられる．

- 実際のタグ付けの精度

  本研究にて学習モデルに用いたタグ付けされたレビューは Khalid ら [1] のタグ付けに基づいて人間がタグ付けした．正解かどうかの検証が不十分で，違ったタグ付けが行われる可能性がある．

表 4 実験 2 の結果：縦軸のカテゴリーのデータがモデルにより横軸のカテゴリーに分類された件数

カテゴリー	A	B	C	D	E	F	G	H	I	J	K	L	正解	不正解	合計	再現率
A. 強制終了	194	18	1	2	11	0	0	0	0	0	6	6	194	44	238	0.81513
B. 互換性	11	258	1	2	37	2	7	3	0	0	9	5	258	77	335	0.77015
C. 機能削除	1	3	24	2	8	0	3	0	0	0	7	0	24	24	48	0.5
D. 機能要求	1	8	4	96	33	0	6	0	0	0	17	0	96	69	165	0.58182
E. 機能エラー	13	28	2	14	421	3	14	7	0	1	36	14	421	132	553	0.7613
F. 隠されたコスト	0	0	0	2	3	13	1	1	0	0	9	0	13	16	29	0.44828
G. インターフェース設計	1	6	1	8	21	0	82	0	0	0	13	4	82	54	136	0.60294
H. ネットワーク問題	0	9	0	1	16	1	0	44	0	0	7	5	44	39	83	0.53012
I. プライバシーと倫理	0	0	2	0	3	0	0	0	7	0	3	0	7	8	15	0.46667
J. 重いリソース	1	3	0	2	4	0	1	1	0	12	3	0	12	15	27	0.44444
K. 魅力のない内容	5	19	0	12	48	2	11	2	0	1	169	2	169	102	271	0.62362
L. アプリが応答しない	4	10	0	1	13	0	1	2	0	0	5	64	64	36	100	0.64
合計	231	362	35	142	618	21	126	60	7	14	284	100	1384	616	2000	
適合率	0.83983	0.71271	0.68571	0.67606	0.68123	0.61905	0.65079	0.73333	1	0.85714	0.59507	0.64				

- 複数のタグの問題

  レビューの中には，内容的に複数のタグ付けがなされるべきものも存在するが，分類方法ではその中の１つのみに分類されることを前提としたが，１つのレビューが複数のタグに当てはまる場合を考えると，各タグにあてはまるかだけを判断する複数の学習モデルが必要となると考えられる．

- 精度向上の問題

  データセットの件数を増加させるとさらに精度が向上するのかを検証する必要があると考える．しかし，実験ではデータセットが増えるにつれて精度の向上の割合が減少していたため，本研究の結果以上に精度が向上しない可能性も考えられる．

- カテゴリー毎のデータ数の問題

  データセットのカテゴリー毎のデータ数を考えると，学習データの 12 種類のカテゴリーのうち，多いカテゴリーでは 500 個を超えるほどのレビューがあるが，少ないカテゴリーは 15 個しかない．カテゴリーごとにデータ数が偏っていることで，予測にばらつきがあることも考えられる．そのため，カテゴリーによっては件数を増やすことにより，精度の向上が期待できる．

- 対象とするアプリのジャンルの問題

  本研究はアプリのジャンルを指定せずに行なったが，例えば，学習モデルに用いたデータセットにゲームのアプリのレビューが多い場合，ゲーム以外のジャンルのアプリを予測した場合は精度が下がってしまう可能性がある．そのため，アプリをジャンルごとや同じ開発会社のアプリごとなどに学習モデルのデータを絞っていくと，精度が向上する可能性も考えられる．

6 まとめと今後の課題

本研究では，過去の研究で収集したレビューをデータセットに用いて，レビューのカテゴリーを予測する学習モデルを作成し，作成したモデルの精度を検証するとともに，どのカテゴリーの予測が正解しやすいのかを調査した．レビューを 2000 件集めて学習モデルを作成して交差検証を行い精度を求めた結果，精度は約 0.68 となった．また，データ数が多いカテゴリーの予測は正解しやすかったが，データ数が少ないカテゴリーの予測は間違えられやすく，各カテゴリーだと判断可能な単語が多いかどうかも予測に影響を及ぼすと考えた．今後，より多くのレビューを追加して学習を行い，モデルの精度の向上を目指すとともに，各カテゴリーの再現率や適合率が向上するのかを検証していく．また，本研究では１つのカテゴリーのみに分類するモデルを作成したが，12 種類のカテゴリーのうち１つのカテゴリーに絞り，そのカテゴリーであるかどうかを判断できるようなモデルを作成してみることで複数のタグに当てはまるレビューに対して複数のタグ付けが可能かどうかや，レビューに含まれる類義語を特定の単語に変換したりレビュー内の高評価に繋がる部分と低評価に繋がる部分を分割してみるなど，レビューの特徴に着目して検証をしていきたい．

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] 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.

[3] 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―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―”，南山大学理工学部 2018 年度卒業論文，2019.

[4] ディープラーニングで文章・テキスト分類を自動化する方法 https://spjai.com/category-classification/

[5] 交差検証（クロスバリデーション）とは？合わせてグリッドサーチに関しても学ぼう！ https://aiacademy.jp/media/?p=263