スマートフォンアプリケーションのレビューの自動分類
―開発会社が同一のアプリを集めた場合の精度向上について―

2020SE060 白木麻衣子　2020SE095 奥野太紀
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（以下，アプリ）のユーザーレビューにはアプリに対する様々な意見や感想が投稿される，それらはアプリの開発者にとって重要な参考意見になる．伊藤ら [2] はアプリの低評価レビューを機械学習により分類する方法を示し，精度向上のための改善策も示したが，その有効性は検証されていない．

本研究では，精度向上のための方策として同じジャンル，さらに作成社が同一のアプリに限定したレビューを学習モデルに用いることで精度の高い機械学習モデルの作成，自動分類ができるかを検証する．実験では，機械学習によってタグ付けされたデータを検証しながら，正しいタグ付けに修正したうえで学習モデルに加えていくことで学習モデルの件数を成長させていき，タグ付けの精度の変化を検証する．十分な精度をもった学習モデルが作成できることで大量のレビューを正しく分析することが可能となり，レビューをアプリの保守・運用に役立てることが期待できる．

2　研究の背景

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

一般的にアプリケーションストア (以下，アプリストア) から配布されるアプリに対して，ユーザーは評価をレビューとして投稿することができる．ユーザーが投稿したレビューは，他のユーザーがアプリを利用する際に参考となる情報であるとともに，アプリの開発者にとっても貴重な情報となり，アプリの運用・保守に役立てられている．

2.2　アプリケーションのレビューを対象とする分析

Khalid らの研究 [1] ではアプリの苦情レビューを精査し，ユーザーがどのような要素に対して不満や関心を抱いているのか調査した．[1] では，レビューの苦情を種類ごとに 12 種類のカテゴリーに分類し，北米で無料で提供されている ios アプリを対象にそれぞれのレビューに当てはまるカテゴリーのタグをつけた．各カテゴリーが低評価レビューの中でどれだけ出現しやすいかを調査した．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付けている．伊藤ら [2] は，レビューのタグ付けの自動化を目的として，Khalid ら [1] のカテゴリーに基づいて機械学習を用いてレビューのカテゴリー分けを行う方法を提案した．実際に 2000 件のレビュー対して交差検証を行った結果，正答率はおよそ 7 割程度となった．[2] では精度向上のための方策として，件数をさらに増やすことや，ジャンルを制限することなどの方法が提言されていた．

2.3　先行研究でのタグ付けのプロセス

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．タグ付けされたレビューについて正しいかどうかの検証を行い，今後の学習データの構築に用いる．

3　作成社が同一のアプリのレビューを集めた場合の分類精度に関する考察

3.1　実験全体の方針と本研究の方針

伊藤ら [2] の研究で機械学習を用いてレビューを苦情の種類ごとに分類する方法の可能性を確認した．レビュー収集の観点から精度の向上を目指す場合，

1. ジャンルを限定せず，レビューの件数を増やす
2. ジャンルを限定する
3. 開発会社を限定する

の 3 つの方策が挙げられる．本研究では「アプリの開発会社を限定してレビューを収集する」方策に基づいて確認をする．アプリの開発会社を限定してすることで，行う施策などに傾向がみられることが期待でき，そのようなレビューをもとに学習モデルを作成することによって，レビューの自動分類の精度が向上するといった仮説のもと，実際に精度が向上するのかを調査する．

本研究では，データセットとして用いるレビューのアプリの作成社を限定することで機械学習によるタグ付けの精度が向上するのかを検証する．機械学習でタグ付けを行った結果を人間の手で分類した結果と比較して精度を確認するとともに，検証したレビューを学習モデルに追加していき学習モデルの構築に用いるレビューの件数を増やす．このように学習モデルの件数を増やすことで，レビューの自動分類の精度がどのように向上するか，単にジャンルを限定した場合と比較して効果が得られるのかを検証する．

4　評価実験

4.1　分析に用いたデータセット

実験では，データセットとして用いるレビューの作成社を限定し，機械学習によるタグ付けの精度が向上するのかを検証する．アプリの作成社は BANDAI NAMCO と SQUARE ENIX の 2 つとし，それぞれレビューを取得する．レビューの取得に使用したアプリと用いたレビューの件数の一部を表 1 に示す．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際に一部のレビューは人間の手によってタグ付けが行われている必要がある．今回は初めに 200 件のレビューを学習用データとして手動でタグ付けし，学習モデルを作成した．タグ付けや評価を行う手順は，先行研究 [2] でのタグ付けのプロセスと同じである．

表 1　使用したアプリと件数の例

BANDAI NAMCO アプリ名	件数	SQUARE ENIX アプリ名	件数
ONE PIECE バウンティラッシュ	300	ドラゴンクエストタクト	300
ドラゴンボールレジェンズ	200	ドラゴンクエストウォーク	200
⋮		⋮	
釣りスピリッツ	100	オクトパストラベラー	200
ドリフトスピリッツ	200	SINoALICE	100
合計	3200	合計	3200

4.2　実験の内容

- 実験 1：学習モデルを用いて新たなレビューに対して予測を行って，予測結果を検証し正しい分類結果にしてから学習モデルに追加する手順を繰り返す．学習モデルの成長に伴って，予測の精度の変化を調査する．
- 実験 2：実験 1 で作成した学習モデルに使用したレビューを用いて学習モデルの成長の各段階において交差検証を行う．学習モデルの成長に伴って，交差検証における予測の精度の変化を調査する．
- 実験 3：実験 1 で作成されたレビューを用いて，作成社ごとに 4 分割交差検証を行い精度を調査する．
- 実験 4：実験 1 で作成した学習モデルに使用したレビューを，もう一方のアプリの作成社の学習モデルを用いてタグ付けを行い，その精度を調査する．

4.2.1　分類するカテゴリーについて

Khalid ら [1] の分類カテゴリーには「強制終了」や「アプリが応答しない」などの似たようなカテゴリが存在する．そのため機械学習でタグ付けを行う際，人によって基準が異なるので精度が落ちるのではないかと考えた．宮下らは [3] で，レビューが抱えている問題の起因に基づいて新しい分類モデルを作成する必要があるのではないかと考えた．[3] で提案された表 2 で示す分類基準を実験で用いた．

表 2　本研究で用いた新たなカテゴリー分け

カテゴリー	カテゴリーの詳細
アプリケーションの問題	アプリケーションの動作自体に問題 (インターフェースに関わることを含む)
会社 (運営) の問題	会社からユーザーによる対応が行われていない
ビジネスモデルの問題	企業のお金の稼ぎ方に関する問題
ユーザーによる問題	ユーザーによるマナーの悪い利用など (システムを介さない人対人のやりとりを含む)
アプリストアの問題	アプリストアが関わってくる問題
コンテンツの問題	ゲームや漫画，その作品自体の問題 (機能的な話は含まない)
ネットワーク，端末の問題	利用者の環境によって起こる問題

4.3　実験の結果

実験 1 の結果を各作成社ごとに図に示す．図の縦軸は正答率，横軸は学習モデルの件数である．図 1 に示す作成社が BANDAI NAMCO では，学習モデルのデータ数が 200 件では 0.50，3100 件では 0.86 の精度が得られ，機械学習を用いたタグ付けの精度が向上していた．学習モデルのデータ数が 1800 件，2000 件では 0.90 とかなり高い精度が得られた．次に SQUARE ENIX の場合では，学習モデルのデータ数が 200 件では 0.55，3100 件では 0.83 とこちらも精度が向上していた．どちらの作成社の場合も実験の後半になるにつれて大幅な精度の向上は見られなかった．

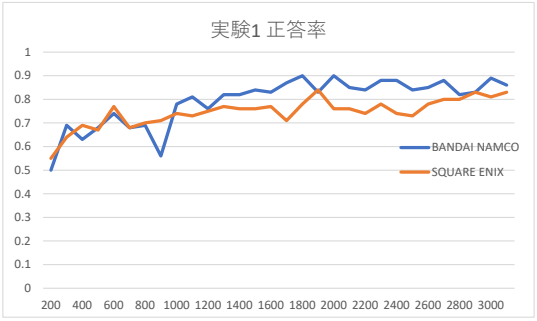

図 1　自動分類における正答率の変化

実験 2 として，各作成社のアプリごとにデータセットに 100 件ずつレビューを追加し，10 分割交差検証を行った結果を作成社ごとに図に示す．図の縦軸は精度，横軸は学習モデルの件数である．図 2 に示す BANDAI NAMCO の場合の学習モデルのデータ数が 100 件の精度はおよそ 0.49，3200 件ではおよそ 0.75 となった．次に SQUARE ENIX の場合の学習モデルのデータ数が 100 件の精度はおよそ 0.53，3200 件ではおよそ 0.76 となり，件数を増やすごとに精度が向上していることが分かった．

実験 3 として，各作成社ごとに 4 分割交差検証を行い集計した，以下では，BANDAI NAMCO 社における結果のみを示す．表 3 を見ると，4 分割交差検証の精度はおよそ 0.76 であり，10 分割交差検証で得られた結果とほぼ同等

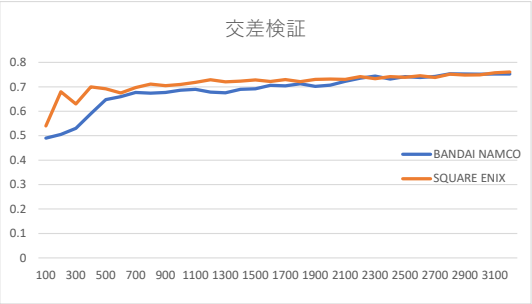

図 2　交差検証における精度の変化

であった．「アプリケーションの問題」とタグ付けされた 832 件のうち 689 件が正しくタグ付けがされ，その再現率が 0.82813 と高い精度であることが分かる．「ユーザーの問題」，「アプリサイトの問題」，「ネットワーク，端末の問題」は件数がそれぞれ 26 件，4 件，20 件と少なく，再現率，適合率ともに高い精度は得られなかった．SQUARE ENIX 社に対しても 4 分割交差検証を行ったところ，精度はおよそ 0.75 であった．

実験 4 の結果を各作成社ごとに表 4 に示す．表 4 は実験 1 で作成された SQUARE ENIX のデータを学習モデルとし，実験 1 で学習モデルに用いた BANDAI NAMCO のレビューを改めて機械学習でタグ付けし直し，その精度，再現率，適合率等をまとめた表である．表 4 を見ると，合計の正解数が 1927 件，不正解数が 1273 件で，およそ 6 割程度の精度となった．学習データが BAMDAI NAMCO でテストデータが SQUARE ENIX の場合でも同様に精度が低下し，およそ 6 割程度の精度であった．

5　考察

5.1　実験 1 の考察

学習モデルを 3100 件で作った際，BANDAI NAMCO では 0.86，SQUARE ENIX では 0.83 という精度となった．しかし，どちらの作成社も正解率が大幅に上下する場面があり，これは学習モデルのデータとテストデータで苦情の傾向が違ったことが原因だと考えられる．また，一部のアプリは 2 回分検証を行ったが，1 度目の検証よりも 2 度目の検証のほうが精度が高くなった．よって，似たような苦情レビューの傾向を持つアプリを集めることが，精度の向上につながると考えられる．自動分類の精度に関してみると，BANDAI NAMCO は学習モデルが 2000 件付近まで，SQUARE ENIX は 3000 件付近までは右肩上がりに精度が向上していたが，それ以降は精度の向上は見られなかったことから，今回の条件で行う実験では 8 割ほどの精度が限界であると考えられる．

5.2　実験 2 の考察

先行研究 [2] では，ジャンルや作成社を限定しているわけではなく，一概に比べられるものではないが，2000 件の学習データに対して 10 分割交差検証をした結果，0.69 の精度が得られた．学習データが 3200 件の際，作成社が BANDAI NAMCO では 0.75，SQUARE ENIX では 0.76 の精度が得られ，先行研究以上の精度を得ることができた．どちらの作成社の場合も学習データの件数が 3000 件を超えてからも少しづつ精度が向上していることから，さらに件数を増やしていくことで精度の向上が期待できる．

5.3　実験 3 の考察

表 3 を見ると，データ件数の多いカテゴリーでは再現率，適合率ともに 7 割以上となったことから，データの件数を増やすことは再現率と適合率の上昇に大きく関連するのではないかと考えられる．このことから，データの少ないカテゴリーのデータの件数を増やすことで，精度はさらに向上すると考えられる．さらなる精度向上策として，2 重のタグの問題がある．表 3 をみると，レビューの内で複数のカテゴリーの内容に言及されていることで他のカテゴリーのタグがつけられたデータの件数が多い．よって，2 重のカテゴリーを内包するレビューを適切のタグ付けできるようにすることがさらに精度を向上させるために必要であると考えられる．

5.4　実験 4 の考察

表 3 と，表 4 を比較すると，全体の精度は 0.76 から 0.60 へと低下している．特に「ビジネスモデルの問題」の再現率は大きく低下しており，これは BANDAI NAMCO ではビジネスモデルの問題が 1 番データ件数が多かったが、SQUARE ENIX ではコンテンツの問題が一番多いといったカテゴリーの傾向の違いから，カテゴリーごとのデータ件数に差が生じてしまったことが原因と考えられる．このことから，精度を向上させるためには似たような傾向をもつレビューをより多く集める必要があると考えられる．

5.5　分類カテゴリーの傾向

どちらの作成社の場合も各カテゴリーの中で「アプリケーションの問題」や「ビジネスモデルの問題」，「コンテンツの問題」はデータ件数が比較的多いが，「ユーザーの問題」や「アプリサイトの問題」，「ネットワーク，端末の問題」には 3200 件中数件ほどしか分類されておらず，各カテゴリーに分類される件数に大きな差が生じてしまっていることが分かる．このことから，今後は学習モデルに用いるデータを収集する際にカテゴリーごとのデータ件数にも着目していくことが必要になると考える．

5.6　他のアプローチとの比較

本研究の結果と，レビューの自動分類の精度を向上させる方策として挙げた以下の 2 つと比較する．

表 3 BANDAI NAMCO アプリの 4 分割交差検証 (1〜3200)

カテゴリー	A	B	C	D	E	F	G	合計	正解	不正解	再現率
A: アプリケーションの問題	689	38	30	0	0	73	2	832	689	143	0.82813
B: 会社 (運営) の問題	62	139	47	1	0	71	0	320	139	181	0.43438
C: ビジネスモデルの問題	25	29	835	0	0	113	0	1002	835	167	0.83333
D: ユーザーの問題	5	6	7	2	0	8	0	28	2	26	0.07143
E: アプリサイトの問題	0	0	0	0	0	4	0	4	0	4	0
F: コンテンツの問題	74	31	101	5	1	778	2	992	778	214	0.78427
G: ネットワーク，端末の問題	8	3	3	0	0	6	2	22	2	20	0.09091
合計	863	246	1023	8	1	1053	6	3200	2445	755	0.76406
適合率	0.80881	0.56504	0.81623	0.25	0	0.73884	0.33333				

表 4 学習データが SQUARE ENIX，テストデータが BANDAI NAMCO

カテゴリー	A	B	C	D	E	F	G	合計	正解	不正解	再現率
A: アプリケーションの問題	602	18	30	0	9	173	0	832	602	230	0.72356
B: 会社 (運営) の問題	70	31	43	0	1	175	0	320	31	289	0.09688
C: ビジネスモデルの問題	56	16	515	0	0	415	0	1002	515	487	0.51397
D: ユーザーの問題	3	1	3	0	0	21	0	28	0	28	0
E: アプリサイトの問題	0	0	0	0	0	4	0	4	0	4	0
F: コンテンツの問題	74	10	126	0	3	779	0	992	779	213	0.78528
G: ネットワーク，端末の問題	11	1	1	0	1	8	0	22	0	22	0
合計	816	77	718	0	14	1575	0	3200	1927	1273	0.60219
適合率	0.737745	0.402597	0.71727	-	0	0.494603	-				

1. ジャンル限定せず大量のレビューを収集した場合 [4]
   本研究では学習モデルが 3200 件の時点で精度はおよそ 7 割後半だった一方，大塚ら [4] の研究では学習モデルが 20000 件の時点で精度はおよそ 6 割前半だった．作成社を限定してレビューを収集することがより自動分類の精度向上に有効であると考えられる．
2. 同ジャンルのアプリからレビューを収集した場合 [5]
   古山ら [5] の研究で実験 3 と同じ内容でジャンルをショッピング系に限定した場合およそ 7 割後半の精度が得られ，ほぼ同じ程度の精度が得られた．しかし，実験 4 においては精度がおよそ 0.49 となり本研究より精度の下がり幅が大きかった．作成社を限定してレビューを収集することは精度向上に有効であると考えられる．

5.7 分類精度向上のための施策

収集するデータの件数をさらに増やすことで，さらなる精度の向上を見込めるのではないかと考える．また，分類カテゴリーごとのデータ件数の偏りを減らすために，カテゴリーごとのデータ件数にも着目してデータを収集することが必要である．また，複数の分類カテゴリーに該当しうるレビューが存在するので，各カテゴリーに当てはまるかを判断する複数の学習モデルが実際の運用では必要になると考える．

6 まとめと今後の課題

本研究では，作成社が同一のアプリに限定して 3200 件のレビューで学習モデルを作成し，より高い精度でレビューの自動分類ができるかを検証した．それぞれ 3200 件のレビューを学習モデルとして交差検証を行った結果，作成社を限定したことで似たような苦情のレビューが多く集まり，機械学習のタグ付けの精度は向上する傾向が得られた．今後の課題として，機械学習のパラメータの調整や，特徴的な単語のみを抽出するなどの分析モデル自体の改良や，複数の分類カテゴリーに該当しうるレビューへの対策などが考えられる．

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] 伊藤陽，紀本光琉：“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―”，南山大学理工学部 2022 年度卒業論文.

[3] 宮下拓也，杉本雄大：“スマートフォンアプリケーションのレビューの自動分類 ―自動分類システムの実現のための分類モデルの構築―”，南山大学理工学部 2023 年度卒業論文.

[4] 大塚冬馬，宇佐美彪雅：“スマートフォンアプリケーションのレビューの自動分類 -大量のレビューを用いた場合の精度向上について-”，南山大学理工学部 2023 年卒業論文.

[5] 古山滉大，柴田晃希：” スマートフォンアプリケーションのレビューの自動分類 -同種のアプリを集めた場合の精度向上について-”，南山大学理工学部 2023 年卒業論文.