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―

2020SE025 古山滉大　2020SE057 柴田晃希
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ユーザーの様々な意見や感想が投稿され，そのアプリの評価を閲覧することができる．それらの投稿は今後アプリを利用するユーザーだけでなく，アプリの開発者にとっても重要な参考意見になる．伊藤ら [1] はアプリの低評価のレビューの分類を自動化することによって大量のレビューを分析の対象にした実用的な学習モデルの作成を行った．精度向上のための方策も示していたが，有効性の確認は行っていない．

本研究では，同じジャンルのアプリのレビューを集めて学習モデルに用いることでより精度の高い機会学習モデルの作成，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について検証し，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．このように学習モデルを成長させていくことでタグ付けの精度がどのように変化するかを交差検証などを用いて検証する．タグ付けを自動化し，高精度な分類器を作成することによって大量のレビューを正しく分類することが可能となり，レビューを分類することでアプリの保守・運用を支援す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におけるレビュー

アプリは Apple App Store や Google Play ストアなどのアプリケーションストア（以下，アプリストア）を介して配布されている．ユーザーは使用したアプリのレビューをアプリストアに投稿できる．投稿されるレビューはタイトル，星評価，アプリに対するコメントで構成される．投稿されたレビューは新しいアプリを探すユーザーにとって参考意見となりアプリ選択の支援に役立つ．さらに，開発者にとってはユーザーからのフィードバックとみなすことができ，アプリの品質向上，ユーザーエクスペリエンスの改善，新機能の提案，セキュリティの確保などに活用でき，今後の方向性を決定するうえで重要な役割を果たす．

2.2　アプリケーションのレビューを対象とした分析

過去にアプリケーションのレビューを対象とした分析は多くなされている．Khalid ら [2] は米国の米国で提供されている iOS の無料アプリを対象に低評価レビューについてどのような内容のレビューが多いのか，どのような種類のコメントが低評価につながるのかを調査した．12 種類の苦情のカテゴリーに分類し，それぞれのレビューにあてはまるカテゴリーのタグをつけた．

伊藤ら [1] は，機械学習によりレビューを分類する方法を提案した．アプリのレビューをカテゴリー分けする際には，Khalid らによる研究 [2] で用いた 12 種類のカテゴリーを利用しタグ付けの作業を行った．機械学習における分類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，または，間違えやすいのかなどの観点から考察した．その結果，学習データが 2000 件のときの交差検証の正解率は，0.68314 になった．精度向上に対する方策として学習データをさらに増やしていくことや，学習させるアプリのジャンルを限定することなどいくつかの提案がされていた．伊藤ら [1] の研究におけるタグ付けのプロセスを説明する．

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，星評価，タイトル，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　同種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　過去の分析における課題と分析の方針

伊藤ら [1] の研究により，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．実際にそのようなシステムを実現するには，更なる精度向上を目的として，どのような方策が有効であるかを検証して確認する必要がある．実際にレビュー収集の観点から精度の向上を目指そうとした場合，以下の 3 つの方策が考えられる．

- 大量にレビューを収集する
  アプリのジャンルを限定せず，大量のレビューを取得し，学習モデルを作成する．
- アプリのジャンルを限定する
  アプリのジャンルを限定し，学習モデルを作成する．

- アプリの開発会社を限定する
  アプリ開発会社を限定し，その会社のレビューのみをデータセットとして学習モデルを作成する．

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ジャンルを限定することで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．このようにして学習モデルに用いるサンプルの件数を増やすことで，機械学習による分類の精度が向上するのか検証する．

4 評価実験

4.1 分析に用いたデータセット

伊藤ら [1] は，ゲームや SNS，ショッピングなど多様なジャンルのアプリのレビューをデータセットとして学習モデルが作成されている．アプリのジャンルが異なれば，使用される単語などレビューの内容も大きく異なるため自動分類する際の精度が下がってしまう可能性があると考えた．そのため，ジャンルを限定することで精度向上に繋がると仮定した．そこで本研究では，動画配信系のアプリとショッピング系のアプリの 2 種類のジャンルを対象とする．それぞれのジャンルに属するアプリのレビューを収集し，それぞれのジャンルに対する学習モデルを構築した．レビューの取得に使用したアプリと学習データの候補となるレビューの件数を表 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必要があるため，一部のレビューはタグ付けが行われていなければならない．本研究では初めに 200 件のレビューを人間の手でタグ付けをし，学習モデルを作成した．追加されたデータに対してタグ付けや評価を行う手順は，先行研究 [1] でのタグ付けのプロセスと同じである．

表 1 本実験で使用したアプリと件数

ショッピング系アプリ	件数	動画配信系アプリ	件数
Amazon	600	Netflix	400
.st	300	U-NEXT	300
SNKRDUNK	300	Amazon Prime Video	300
Adidas	200	Hulu	300
MUJIpassport	200	AbemaTV	300
NIKE	200	YouTube	300
ユニクロ	200	DAZN	300
HandM	200	Lemino	300
ZARA	200	Tver	200
Wish	200	NHK+	100
GU	200	WOWOW	100
NIKESNIKER	100	Disney+	100
Qoo10	100	Twitch	100
creema	100	FOD	100
Rakuten shopping	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．このようにして学習モデルが成長していくにつれて，予測の精度がどのように変化するのかを調査する．
- 実験 2：実験 1 で作成した学習モデルの構築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 10 分割交差検証を行う．このようにして学習モデルが成長していくにつれ，交差検証の精度がどのように変化するのかを調査する．
- 実験 3：実験 1 で作成した学習モデルに対し，ジャンルごとに 4 分割交差検証を行い，交差検証における予測の精度を調査する．手動でタグ付けしたレビューの各カテゴリーの件数と機械学習によって自動分類されたレビューの各カテゴリーの件数を集計する．集計した結果から各カテゴリーの正解率，不正解率，適合率，再現率を求める．

4.3 分類するカテゴリーについて

Khalid ら [2] が用いたカテゴリーでは，類似しているカテゴリーが存在するため，人間がタグ付けを行う際に，人によって分類結果が異なる可能性がある．そこで宮下ら [3] は，レビューが抱えている問題が何に起因しているかに基づいて新しい分類モデルを作成した．実験では表 2 に示す [3] で提案された新たな分類基準を用いる．

表 2 本研究で使用する新たなカテゴリー

——カテゴリー——	——カテゴリーの詳細——
アプリケーションの問題	アプリの動作自体に問題がある
会社（運営）の問題	会社からユーザーに対す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	ユーザーの利用マナーなどの問題
アプリストアの問題	アプリのストアに関する問題
コンテンツの問題	アプリ内のコンテンツに関する問題
ネットワーク, 端末の問題	ユーザーの利用環境に関する問題

4.4 実験 1

ショッピング系アプリと動画配信系アプリのそれぞれで，はじめに手動でタグ付けを行った 200 件のレビューを初期学習モデルとした．タグ付けが行われていないレビューを 100 件ずつ機械学習を用いてタグ付けを行い，分類精度を調査したうえで正しいタグ付けに修正して学習モデルに追加した．この操作を繰り返して，各段階での学習モデルが新しい 100 件のレビューに対してどれだけ正しくタグ付けできたかを図 1 に示す．縦軸が正解率，横軸がレビューの件数を表す．ここの正解率は，機械学習で判断した結果が人の手でタグ付けした結果と等しかった割合を示す．図 1 からショッピング系のアプリでは，学習モデルのデータ数を 400 件時点までデータを増やすと正解率が上昇する傾向がみられ，それ以降は精度にばらつきがみられ

た．動画配信系のアプリでは，ショッピング系のアプリと同様に学習モデルのデータ数を 400 件時点までデータを増やすと正解率が上昇する傾向がみられたが，それ以降は正解率の上昇割合が減少した．全体としては学習モデルの件数の増加とともに正解率が上昇する傾向にあった．

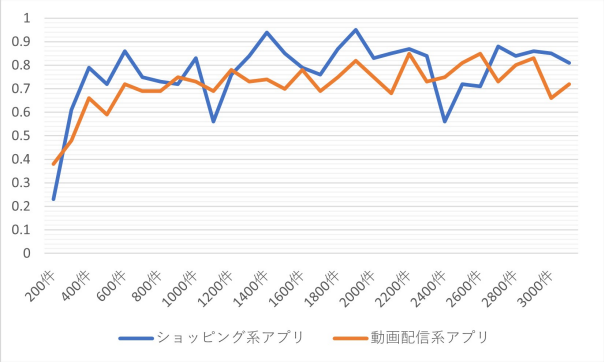

図 1　追加したデータに対する分類結果の正解率の推移

4.5　実験 2

ショッピング系のアプリと動画配信系のアプリそれぞれで，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．学習モデル内のレビューの件数が 200 件から順次 10 分割交差検証を行った結果を図 2 に示す．縦軸が正解率，横軸がレビューの件数を表す．学習データが 3200 件の際の精度はショッピング系のアプリで 0.7918 ，動画配信系のアプリで 0.7300 となった．ショッピング系のアプリは，図 2 に示すように学習モデルのデータ数が 3200 件未満ではデータ数を増やしても正解率の上昇率が不規則であることがわかった．動画配信系のアプリは，学習モデルのデータ数を増加に伴い，正解率も上昇する傾向がみられた．

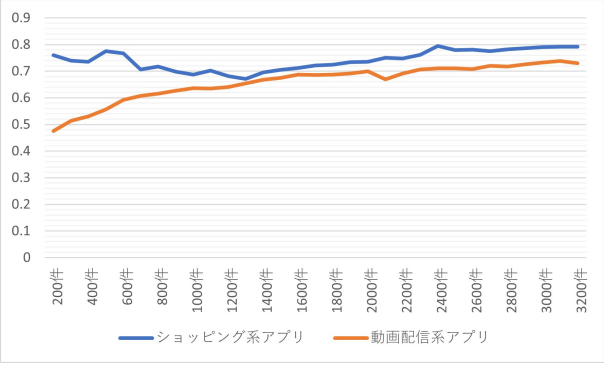

図 2　交差検証における正解率の推移

4.6　実験 3

ショッピング系のアプリと動画配信系アプリのそれぞれで，3200 件のデータセットを対象に，4 分割交差検証を行った．ショッピング系のアプリの結果を表 3，動画配信系アプリの結果を表 4 に示す．ショッピング系のアプリでは，合計時の再現率が 0.79688 となった．動画配信系アプリでは，合計時の再現率が 0.7403125 となった．

5　考察

アプリのジャンルをショッピング系のアプリと動画配信系のアプリの 2 種類に絞ってそれぞれのデータセットに対して分析を行ったところ，実験 1 では，両ジャンルともに，内容が類似しているレビューが多いアプリの正解率は，正解率が高くなる傾向がみられた．それは，分類されるカテゴリーが一部に偏るためだと考える．また，実験 2 ではジャンルごとに特徴が異なる正解率の推移が得られた．

5.1　ショッピング系アプリの考察

ショッピング系のアプリでは実験 2 において，学習モデルのデータ数が 200 件時点から正解率が 0.7600 と高い数値が得られ，その後も比較的高い正解率であったものの，上昇率にはばらつきがあった．学習モデルに追加されたレビューのカテゴリーに偏りがあったためだと考えられる．学習モデルにおいて最初の 200 件のうち 94 件がアプリケーションの問題，88 件が会社 (運営) の問題に該当するものであり，レビューのカテゴリーに大きな偏りがあったため最初から高い正解率が得られたと考える．最終的な学習モデル全体でも 3200 件中 2685 件がアプリケーションの問題及び会社 (運営) の問題に該当するレビューであったことで，全体的に正解率が高くなったと考える．また，学習モデルにレビューを追加していくにつれて，他のカテゴリーに該当するものが徐々に増えていったことで正解率の上昇率にばらつきがみられたと考察する．今後さらにデータセットの件数を増やしていき，分類精度の向上が見込めるか検証していく必要があると考える．

5.2　動画配信系アプリの考察

動画配信系のアプリにおいて，実験 1 で正解率にばらつきがでた．同じ動画配信系のアプリでも広告の有無や課金制か否かによってレビューの内容の傾向が大きく異なる傾向があったからだと考える．実験 2 では図 2 に示すように，学習モデルのデータ数が増加するほど正解率が上昇したが，データ数が増加するほど正解率の上昇率が減少している．これらの結果から，分類精度の向上のためにさらなるデータの取得が必要であると考えられる．実験 3 では，表 4 からわかるように「アプリケーションの問題」や「会社 (運営) の問題」など，データ数の多いカテゴリーほど再現率，適合率が高い傾向があるため，他のカテゴリーと比べてデータ数の少ないカテゴリーのレビューを増やし，学習させることで全体の分類精度向上につながると考える．

5.3　他方のデータを分類した場合の結果

ショッピング系アプリと動画配信系アプリで学習モデルを交換して，学習モデルが 3200 件のレビューに対してどれだけ正しくタグ付けできたか検証した．その結果，合計時の再現率が 学習モデルがショッピング系アプリでテス

表 3 ショッピングアプリの 4 分割交差検証 (合計)

カテゴリー	A	B	C	D	E	F	G	正解	不正解	合計	再現率
A: アプリケーションの問題	1576	126	6	2	0	23	0	1576	157	1733	0.90941
B: 会社 (運営) の問題	179	690	21	22	0	40	0	690	262	952	0.72479
C: ビジネスモデルの問題	26	41	36	0	0	2	0	36	69	105	0.34286
D: ユーザーによる問題	7	27	5	40	0	3	0	40	42	82	0.4878
E: アプリストアの問題	3	1	0	0	0	0	0	0	4	4	0
F: コンテンツの問題	61	41	1	2	0	208	0	208	105	313	0.66454
G: ネットワーク・デバイスの問題	8	2	0	0	0	1	0	0	11	11	0
合計	1860	928	69	66	0	277	0	2550	650	3200	0.79688
適合率	0.84731	0.74353	0.52174	0.60606	-	0.7509	-				

表 4 動画配信系アプリの 4 分割交差検証 (合計)

カテゴリー	A	B	C	D	E	F	G	正解	不正解	合計	再現率
A: アプリケーションの問題	1230	113	56	0	1	13	44	1230	227	1457	0.8442
B: 会社 (運営) の問題	166	314	44	0	1	16	28	314	255	569	0.55185
C: ビジネスモデルの問題	65	46	602	0	0	8	8	602	127	729	0.82579
D: ユーザーによる問題	2	1	0	4	0	0	0	4	3	7	0.57143
E: アプリストアの問題	6	0	0	0	0	0	0	0	6	6	0
F: コンテンツの問題	46	33	26	0	0	61	4	61	109	170	0.35882
G: ネットワーク・デバイスの問題	70	20	11	0	0	3	158	158	104	262	0.60305
合計	1585	527	739	4	2	101	242	2369	831	3200	0.74031
適合率	0.77603	0.59583	0.81461	1	0	0.60396	0.65289				

トデータが動画配信系アプリの場合は 0.488125，学習モデルが動画配信系アプリでテストデータがショッピング系アプリの場合は 0.5719 となった．同じカテゴリーと分類基準を用いても，学習モデルとテストデータのアプリのジャンルが異なると再現率が大きく低下することがわかる．

5.4 他のアプローチとの比較

大塚らの研究 [4] ではアプリをジャンルを限定せず，大量のレビューを収集して機械学習を行うことで分類精度がどのように推移するかの検証を行った．交差検証の結果，データ数が 3200 件の時点で正解率が 0.60437 であった．本研究と比較してわかるようにジャンルを限定せず収集したレビューの学習モデルよりも，本研究のジャンルを限定したレビューの学習モデルの方が正解率が高い値となった．この結果からアプリのジャンルを限定することによってレビューの内容のばらつきを減らすことができ，分類精度が比較的高くなることが考えられる．

白木らの研究 [5] ではアプリの開発会社を限定してレビューを収集し，機械学習を行うことで分類精度がどのように推移するのかの検証を行った．データ数が 3200 件の時点で正解率がそれぞれ 0.7525 と 0.7606 であった．比較すると本研究の動画配信系アプリの結果よりは 2 つとも最終的な正解率が高くなっているが，ショッピング系アプリの結果よりは低くなっており，ジャンルを限定することが精度向上に有効であることがわかる．学習モデルとテストデータを交換した場合の結果を比較すると，ジャンルが近いデータ同士では精度が下がりにくいことや組織における傾向も精度の向上に重要であると考えられる．

6 まとめと今後の課題

本研究では，ショッピングアプリと動画配信系アプリのレビューのカテゴリーを予測する学習モデルをそれぞれ作成し，アプリのジャンルを限定することで分類精度が向上するかを検証した．ジャンルを限定せず，大量のレビューを収集した場合と比較して，精度が全体的に改善しており，ジャンルを限定することで精度が向上することわかった．今後の課題として，複数タグに対応した分類環境の実現，機械学習のパラメータの調整や，特徴的な単語のみを分類に用いるなどの方策の有効性を確認したいと考えている．

参考文献

[1] 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.

[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] 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，2024.

[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 年度卒業論文，2024.

[5] 白木麻衣子，輿野太紀：“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―”，南山大学理工学部 2023 年度卒業論文，2024.