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―

2020SE048 大塚冬馬　2020SE072 宇佐美彪雅
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) におけるユーザレビューは，アプリのユーザによる様々な意見が多く投稿されており，アプリ利用者にとって参考になる意見であるとともに，アプリ開発者にとっては今後のアプリの保守や運用の際の重要な情報源である．伊藤ら [2] はアプリの低評価レビューを自動で分類する学習モデルを作成し，苦情内容に基づいて機械学習で分類を行う方法を作成した．今後の課題として分類精度の向上を挙げており，具体的な方策としてより多くのレビューを追加して学習を行うことを提案していたが，効果について検証していない．

本研究では，データセットとして用いるレビューの件数を増加させることでタグ付けの精度が向上するのかを検証する．実験ではジャンルを絞らずレビューを集め，学習モデルを作成し，タグ付けを行う．機械学習によってタグ付けされたレビューを手動で正しいタグに付け直し学習モデルに追加していくことで件数を増やしながら精度がどのように向上するかを分析するとともに，交差検証を行い評価を行う．高い精度の学習モデルが作成できることでタグ付けを自動化でき大量のレビューを正しく分類することが可能となり，レビューを今後のアプリの保守・運用に効率的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におけるレビュー

基本的にアプリは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．アプリをダウンロードしたユーザはそのアプリの評価をアプリストアに投稿することが出来る．レビューは，主にタイトル，星評価，アプリに対するコメントの三つから構成される．ユーザにより投稿されたレビューは，他のユーザの参考意見になるとともにアプリ開発者にとっても貴重な参考意見ともなっている．アプリ開発におけるユーザレビューは，アプリの保守や運用の際の重要な情報源と考えられる．

2.2　アプリケーションのレビューを対象とした分析

アプリのレビューを対象とした分析は過去に多くされてきた．Khalid ら [1] は，北米で提供されている無料 iOS アプリを対象に，レビューを苦情の種類ごとに 12 種類のカテゴリーに分類した．それぞれのレビューに対して当てはまるカテゴリーのタグを付け，どのカテゴリーの苦情が多いか，どのカテゴリーが低評価に結び付きやすいかを調査した．調査結果では，「機能要求」，「機能エラー」，「強制終了」が苦情頻度の高いカテゴリーとして挙げられ，「プライバシーと倫理」，「隠されたコスト」が低評価の頻度が高いカテゴリーとして挙げられた．これらの情報はアプリを改善する際のリソース配分時に役立つと結論づけている．

伊藤ら [2] は，アプリのレビューをカテゴリーに分類する際に Khalid ら [1] で用いた 12 種類のカテゴリーに基づいて，レビューのタグ付けを自動化することを目的として，機械学習によってタグ付けを自動化する学習モデルを作成した．タグ付けされた 2000 件のレビューに対して交差検証を行った．調査結果では，2000 件に対する交差検証の精度が 0.68314 という結果を得た．精度向上のアプローチとして，レビューの件数をさらに増やしていくことやジャンルを制限することなどいくつかの方法が提案された．

2.3　先行研究 [2] におけるタグ付けのプロセス

先行研究 [2] でのタグ付けのプロセスを示す．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．

- 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．タグ付けした結果について正しいかどうかの検証を行う．正しく修正したうえで，今後の学習データの構築にも用いる．

3　大量のレビューを対象とした場合の分類精度に関する調査

3.1　過去の分析における課題

先行研究 [2] では，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を確認した．実際にそのようなシステムを実用的に活用できるようにしていくためには，更なる精度の向上を目的として，どのような方策が有効であるかを調査する必要がある．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 つが考えられる．

- データセットを増加させる
  アプリのジャンルを限定せず，大量のレビューを取得

し，学習させる．データセットの件数を増加させることで分類精度がどのように変化するかを検証する．

- アプリのジャンルを限定する
  ジャンルを限定して収集したレビューをデータセットとして学習モデルを作成する．ジャンルを限定することで分類精度がどのように変化するか検証する．
- アプリの開発者を限定する
  開発会社を限定し，その会社のアプリ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．データセットを同じ会社のアプリに限定することで分類精度がどのように変化するかを検証する．

3.2 本研究における分析方針

本研究では，データセットをより増加することに着目して，大量のレビューを機械学習させることでタグ付けの精度が向上するのかを検証する．実験ではジャンルを絞らずレビューを集め，学習モデルを作成し，タグ付けを行う．機械学習によってタグ付けされたレビューを手動で正しいタグに付け直して学習モデルに追加し，件数を増やしながら精度がどのように向上するかを分析する．さらに，4 分割交差検証を行って分類カテゴリの傾向や分類の再現率と適合率を求め，結果から分類精度を調査する．また，担当以外の 2 つの方針の結果と比較し，どのアプローチが精度向上に効果的であるかを比較する．

3.3 分析の仮説

本研究で行う分析ではレビューを大量に増やすという施策を用いて行う．この施策で期待できる仮説としては，レビューを増やすことで単純に機械学習に学ばせられる対象が増え，あらゆるジャンルやカテゴリのレビューに対応できるようになるということや，様々な単語や助詞などの言葉がどのような場面で使われるかを学習することで分類精度が向上するという効果が得られると考えられる．

3.4 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「重いリソース」,「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．そのため機械学習によるタグ付けを行う際，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．そこで，宮下らは [3] レビューが抱える問題が何に起因するかに基づいた新しい分類モデルを作成する必要があるではないかと考え，分類モデルを提案した．表 1 に，[3] で提案されている新しい分類基準を示す．

表 1 宮下ら [3] で提案されている分類基準

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 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件数を多くすることでタグ付けの精度が向上するのかを検証する．ジャンルは指定せずにレビューを収集し，学習モデルを作成した．レビューの取得に使用した 81 個のアプリとそのレビューの件数の一部を表 2 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，一部のレビューは手動でタグ付けをしておく必要がある．本研究では伊藤ら [2] と同じ手順で分析を行った．分析では，初めに 100 件のレビューをデータセットとして手動でタグ付けをし，学習モデルを作成した．

表 2 使用したアプリと件数

アプリ名	件数	アプリ名	件数
Twitter	400	Threads	400
妖怪ウォッチぷにぷに	100	パズドラ	500
Line	400	KaKao Talk	300
Instagram	400	Yahoo Japn	400
Linemusic	300	・・・	・・・
Gmail	400	合計	20000

4.2 実験の手順

本研究では以下の 3 種類の実験を行う．

- 実験 1 初期の学習モデルに対して，レビューを 100 件づつ判定を行いながら正しいカテゴリに修正し，学習モデルに追加していくという操作をする．追加したレビューの件数が増えていく過程で，学習モデルの判定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対して交差検証を行う．レビューの件数が増えていく過程で，交差検証の精度の変化を調査する．
- 実験 3 実験 1 で用いたデータセット 20000 件を 4 等分してそれぞれ 5000 件のレビューが含まれるデータ A，データ B，データ C，データ D として，そのうち 3 つのデータをデータセットとするモデルを 4 種類作成し，それぞれ残った 1 つのデータを分類対象としてカテゴリーを予測させ，カテゴリーごとに正解数や不正解数，再現率や適合率を求める．

4.3 実験 1 の説明

実験 1 では 100 件〜2000 件まで 100 件ずつデータセットを追加し 2000 件〜5000 件までは 200 件ずつデータセットを追加する．5000 件〜20000 件までは 500 件ずつデー

タセットを追加した．学習モデルが成長していく過程の機械学習によるタグ付けの精度の変化を図 1 に示す．結果を見てみると，データセットが少ない 100〜5000 までは，一番精度が高い時は 0.95 なのに対して一番精度が低い時は 0.26 という結果だった．精度が大幅に上がったり，下がったりしていて分類の精度が安定していないのが分かる．データセットが 5500〜20000 では，精度が上がったり下がったりしているが大幅に精度が変わる頻度は少なく緩やかに精度が向上した．しかし，正答率が 0.8 以上を超えることはなかった．

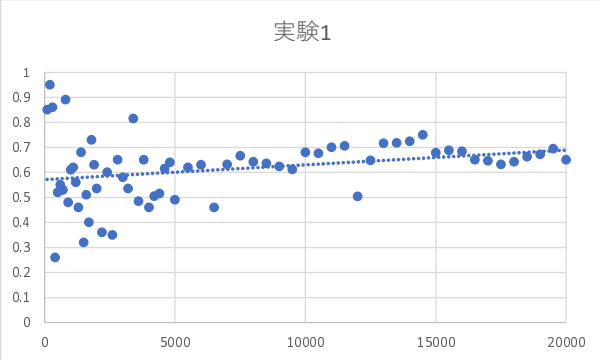

図 1　追加したデータに対する分類結果の正解率の推移

4.4　実験 2 の説明

実験 2 では実験 1 で作成した学習モデルが成長していくにつれ交差検証を行った．学習モデルが成長していく過程で，交差検証の精度の変化を図 2 に示す．分類結果を見てみると，データセットが少ない 100〜5000 までは，一番精度が高い時は 0.648 なのに対して，一番精度が低い時は 0.492 という結果だった．精度が大幅に上がったり，下がったりするわけではないが分類の精度が安定して向上しているわけではなく緩やかに向上した．また 100 件〜2000 件までの精度の平均は 0.586 だった．学習データが 5500〜20000 では，一番精度が高い時は 0.694 なのに対して，一番精度が低い時は 0.578 と精度が上がったり下がったりしているが大幅に精度が変わることはなかった．また 5500 件から 20000 件までの正答率の平均は 0.625 で 100 件〜2000 件までの精度の平均より高くなった．しかし，正解率が 0.7 以上を超えることはなかった．

4.5　実験 3 の説明

実験 3 では集めた 20000 件のレビューの 4 分割交差検証を行った．まとめた結果を表 3 に示す．20000 件のうち，最も再現率が高いのは「会社の問題」の 0.628 であり，最も再現率が低いのは「アプリストアの問題」の 0.284 であった．また，最も適合率が高いのは「会社の問題」の 0.714 であり，最も適合率が低いのは「ネットワーク・デバイスの問題」の 0.39 となった．

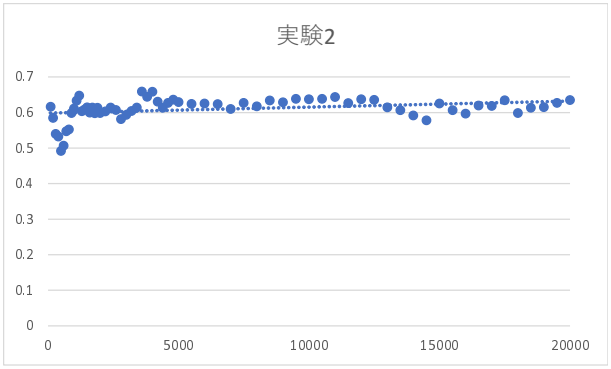

図 2　交差検証における正解率の推移

5　考察

実験 1 では，図 1 の近似曲線から分かるように分類精度は緩やかに上昇したが，大幅な精度向上は見られなかった．しかし，100 件〜 5000 件までの精度のバラつきが件数を増やしていくに連れて落ち着いて行っていることから，このまま件数を増やしていくことによって安定した分類精度の上昇が見られることが考えられる．件数が少ない時の精度のバラつきは学習モデルに含まれる件数が少ないジャンルのアプリを分類した時に多く見られたため，学習モデルに含まれる件数が増えてジャンルごとのレビューが増加したことでバラつきが収まったと考えられる．20000 件ではそれぞれのジャンルの件数をある程度まで増やすまでしか至らなかったため，今回の研究ではまだデータ数不足であると感じた．これらより 20000 件よりもさらに大幅にデータ数を増やすことで精度向上させることが期待できる．

実験 2 では，学習モデルを 100 件で作成した際は正解率が 0.616 ，20000 件で作成した際は 0.635 となり，最終的に精度はやや上昇した．しかし，件数を増加させていく過程で精度は多少の上下を繰り返しており，大幅な精度向上は見られなかった．実験結果を見ると，精度が上昇している際は学習モデルに多く含まれている同一のアプリや，学習モデルに含まれる件数の多いジャンルのアプリのレビューを学習モデルに追加した時に多く見られた．また，精度が下がっているのは学習モデルに含まれていないアプリや，学習モデルに含まれる件数の少ないジャンルのアプリのレビューを追加した時に見られることが多かった．実験 2 の結果から，分類精度を向上させるにはそれぞれのジャンルで，より多くのアプリからレビューを集めて学習モデルの件数の大幅な増加が必要であると考えられる．また，実験 2 の結果より同じアプリやジャンルを学習モデルに多く含まれると精度が上昇し，学習モデルに少ないまたは初めて追加されるアプリを追加すると精度が下がる傾向にあることが分かった．よって件数増加に加えジャンルを限定することで精度の上昇を見込めることが考えられる．

実験 3 では，表 4 に実験 3 の表 3 の再現率から見たカテ

表 3　4 分割交差検証を行った結果

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	2726	936	172	4	10	533	65	2726	1722	4448	0.613
B: 会社の問題	1493	5357	292	18	12	1232	126	5357	3173	8530	0.628
C: ビジネスモデルの問題	99	349	1181	3	2	361	18	1181	832	2013	0.587
D: ユーザーの問題	20	29	10	88	0	70	1	88	130	218	0.404
E: アプリストアの問題	16	21	2	0	27	26	3	27	68	95	0.284
F: コンテンツの問題	492	673	277	13	3	2758	26	2758	1484	4242	0.65
G: ネットワーク・デバイスの問題	86	141	12	0	1	61	153	153	301	454	0.337
合計	4934	7506	1946	126	55	5041	392	12290	7710	20000	0.615
不正解	2208	2149	765	38	28	2283	239				
適合率	0.552	0.714	0.607	0.698	0.491	0.547	0.39				

表 4　再現率から見た合計の割合

カテゴリー	A	B	C	D	E	F	G	合計
A.アプリケーションの問題	0.613	0.21	0.039	0.001	0.002	0.012	0.015	4448
B.会社(運営)の問題	0.155	0.628	0.034	0.002	0.001	0.144	0.015	8530
C.ビジネスモデルの問題	0.049	0.173	0.587	0.001	0.001	0.179	0.009	2013
D.ユーザによる問題	0.092	0.133	0.046	0.404	0	0.321	0.005	218
E.アプリサイトの問題	0.168	0.221	0.021	0	0.284	0.274	0.032	95
F.コンテンツの問題	0.116	0.159	0.065	0.003	0.001	0.65	0.006	4242
G.ネットワーク、端末の問題	0.189	0.311	0.026	0	0.002	0.134	0.337	454

ゴリー別の割合を示す．表についている色は数値が高いほど濃く，低いほど薄くするように設定した．この二つの表から，再現率の観点から見てすべてのカテゴリにおいて正しく分類された割合が最も高くなったことが分かった．適合率の観点からも同様の結果が得られた．表 3 から，それぞれのカテゴリにおける再現率と適合率を見ると，再現率が 0.6 以上と高かったカテゴリは「コンテンツの問題」の 0.65 と，「会社の問題」の 0.628,「アプリケーションの問題」の 0.613 であった．これらは全カテゴリにおいて学習モデルに含まれるレビュー数が多いカテゴリである．よって再現率の高さは学習モデルに含まれるレビュー数に関係すると考えられる．他のカテゴリにおいても再現率の高さは学習モデルに含まれるレビュー数と比例していることから，再現率を向上させるためには学習モデルに件数の少ないカテゴリのレビューを増やすことが必要であると考えられる．また，適合率に着目しても数値が最も高くなったのは再現率と同じく「会社の問題」であった．しかし，次に高くなったのは分類されたレビュー件数の少ない「ユーザーの問題」となった．このように適合率はレビューの件数による影響もあるが，カテゴリの分類のしやすさにも起因するということも考えられる．「ユーザーの問題」はレビュー数が少ないにも関わらず適合率が高いことから，正しく分類されやすいことが考えられる．よって適合率を向上させるためにはレビュー数を増やすという施策と，カテゴリごとに分類しやすい単語を残したり不要な単語を除去することで分類しやすくするという施策が考えられる．

6　まとめと今後の課題

本研究では大量のレビューを収集して学習モデルを作成し，機械学習による自動分類の精度が向上するか検証した．データセットを増やしていくにつれ精度が減少したり上昇したりしているが緩やかに向上した．

本研究の「データセットを増加させる」という方針で実用化できるほどの分類精度にするにはより長い時間をかけてレビューの件数を増やしていく必要がある．今後，レビュー内容の特徴的な単語を残したり，不要な単語や文章を除去するなど，機械学習に与えるデータを加工したり，学習におけるパラメータの調整などを行って，分類の精度がどのように変化するかを検証していきたい．

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] 伊藤陽，紀元光琉：“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，2023.

[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.