クラウドソーシングと自然言語処理による安価な UX 評価の実現

小川健太郎　奥川真理子　加藤晃子
ヤフー株式会社
{keogawa, mokugawa, akkato}@yahoo-corp. jp

概要

自然言語処理の技術革新が加速している一方で，自然言語処理技術のサービス適用には，専門知識を持った技術者と，高い処理能力を有する計算機資源が必要になる．そのため「導入したくても導入できない」，「導入しても自分たちで運用できない」という企業や組織も多いのが実情である．

我々はこの問題をクラウドソーシングと既存ツールのシンプルな組み合わせで解決した．本稿では「AI リテラシの高くない組織が簡便かつ安価に自然言語処理技術を導入し業務効率化を達成」した事例とその実現プロセスを紹介する．

1 背景

100 以上のサービスをもつヤフーでは，サービス品質の維持・向上への取り組みが必要となるため，品質評価部門が自社の「Yahoo!クラウドソーシング」[1] を利用して UX（User Experience）評価のスキームを構築した．本章では，この UX 評価の概要と，運用上の課題について述べる．

1.1 UX 評価とは

UX 評価は Yahoo!クラウドソーシングで実施するアンケートがベースとなる．実施の流れは，クラウドソーシングに UX 評価のアンケートを入稿するところから始まる．アンケートはサービスの使い勝手や，信頼性などを選択肢から選ぶ選択式の設問と，サービスの良い点や改善すべき点を自由にコメントする自由記述式の設問（図 1）の大きく 2 つある．

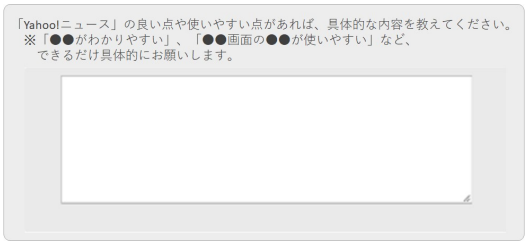


図 1 自由記述式の設問

アンケートは評価対象となるサービスを実際に利用しているユーザーにメールで配信される．

アンケート終了後，品質評価部門において結果を集計し，考察とともにサービス担当者にフィードバックされる．なお，自由記述式の回答については，後述する「UX ピラミッド」を用いて評価する．

1.2 UX ピラミッドとは

UX ピラミッドは 1940 年に提唱されたアブラハム・マズローの欲求階層説を基にアーロン・ウォルターがヒエラルキーを厳密に反映したユーザーニーズのヒエラルキーについて説明したものである[2].

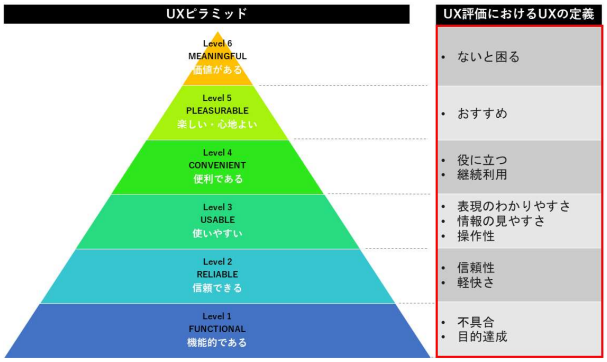


図 2 UX ピラミッドと UX の定義

UX ピラミッドでは，より基本的なニーズ（機能性や使いやすさなど）が満たされた後にのみ，優れたニーズ（ピラミッドの最上部にある喜びや価値など）を達成できるとしている(図 2).

ピラミッドのレベル 1 から 3 は，目的のタスクを達成するユーザーの能力，レベル 4 から 6 では，サービスを使用する際のユーザーエクスペリエンスに焦点を当てており，一般的にユーザー体験のクオリティを測る方法として用いられている．

品質評価部門ではこのレベルをさらに 11 個の項目に細分化し，クラウドソーシングで収集したコメントと紐づけて傾向を分析．これにより，サービスの強みと課題を明らかにし，サービス品質向上のために役立てられている．

1.3 現状の課題

UX 評価の運用において，品質評価部門では大きく以下２つの課題を抱えていた．

課題１：人手でコメントを分類するのに多大な工数が掛かっている

課題２：自動化には自然言語処理に対する専門知識と高スペックな開発環境が必要

「課題１」に関しては，現状，品質評価部門が手作業でコメントを分類しているため 1 サービスにつき 2 日（約 16 時間）の作業工数が掛かっている．

「課題２」に関しては，記載の通り，品質評価部門に専門知識を有する技術者はおらず，分類を自動化するツールの開発は困難な状況であった．

そこで，品質評価部門は，ヤフー社内でデータサイエンスの導入支援を担う我々のチームと連携し，解決策を検討していくことになった．

解決策の策定にあたり，あらかじめ両者で以下の要件を満たすことを条件とした．

要件１：現状より分類に掛かる工数が減ること

要件２：素人でも機械学習モデルの利用がしやすいライトな枠組みであること

2 提案手法

まず，我々が着目したのは，これまでの UX 評価の運用で既に「手作業で分類したデータ」が蓄積されていたこと．我々はこれを学習データとしてコメント分類モデルの構築を試みた．

2.1 学習データの特徴

クラウドソーシングで得られるコメントは「端的なコメントが多い」といった特徴があげられる．

例えば，Yahoo!ニュースの場合は以下のようなコメントが寄せられることが多い．

例）・見出しがわかりやすい
・ニュースの更新頻度が高くて良い
・ユーザーが投稿するコメントが楽しい

図 3 は各アプリマーケットとクラウドソーシングに投稿された Yahoo!ニュースアプリに対するレビューコメントの中から直近の投稿 200 件を抽出し文字数の分布を可視化した結果である．コメントの平均文字数は App Store(Apple)は 54 文字，Google Play は 47 文字であるのに対し，クラウドソーシングは 20 文字と全体的に少ない傾向であった．なお，集計するうえで「不真面目な回答」[3] は除外している．

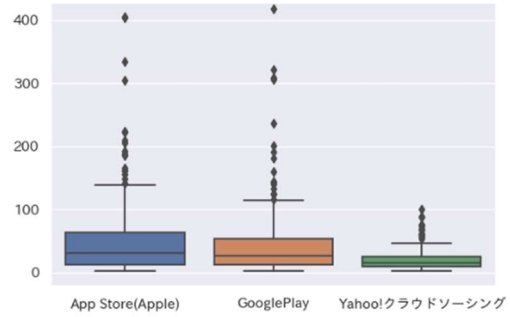


図 3 レビューの文字数（Yahoo!ニュース）

2.2 fastText の利用

コメント分類モデル構築においては，「ライトな枠組み」という要件を鑑み，主に導入容易性の観点から自然言語処理ライブラリ「fastText」[4] を使用した．

2.3 コメント分類の流れ

今回は図 4 で示す 5 つのプロセスでコメント分類モデルを構築した．以下，順を追って説明する．

図 4 コメント分類モデルの構築の流れ

① 手作業でのコメント分類

前述の通り，既に「手作業で分類したデータ」が手元にある状態であったが，内容を見るとかなり不均衡なデータであった（図 5）．特に「08. 継続利用」や「11. ないと困る」に分類されるコメントが少なかった．これらを補うため，品質評価部門とともに新たなコメントを作成し，学習データに追加した．最終的に 5, 426 件の「手作業で分類したデータ」が準備できた．

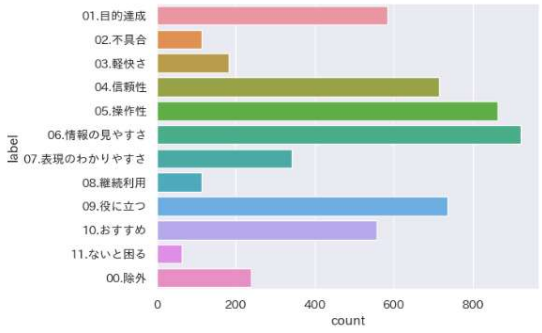


図 5 コメント分類ごとの分布

② 学習データの形式に置換

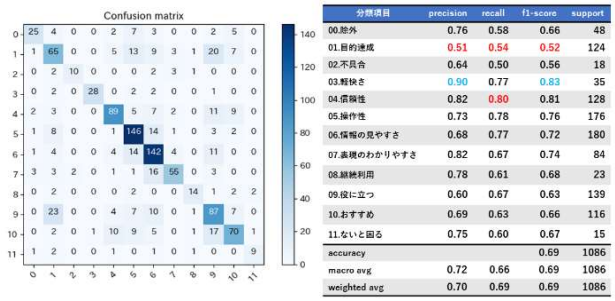
「手作業で分類したデータ」を fastText の学習データの形式に置換する．具体的には項目名を先頭に配置し「__label__」というプレフィックスを付与するのみであるため，専門知識を持たない品質評価部門も容易に学習データを作成することができた．

③ 素性抽出

形態素解析エンジン「MeCab」[5] を利用して，コメントから素性を抽出．当初「動詞」「形容詞」「名詞」の 3 品詞の形態素のみを対象としていたが，「使いやすい」で「使い」のみが，「○○できない」で「○○」と「でき」のみが抽出されてしまうなど，ユーザーの感じ方が得られにくい状況となっていた．解決策として「接尾辞」も対象とし，「使い　**やすい**」「○○　でき　**ない**」といったニュアンスを抜き出せるよう工夫した．

④ モデリングと評価

上述の通り，自然言語処理ライブラリ「fastText」と手作業で分類した 5, 426 件のデータを用いてコメント分類モデルを作成．データのうち 8 割（4, 340 件）を学習に，2 割（1, 086 件）を評価に利用した．今回はクロスバリデーションを実施せず，不均衡な学習データと同等の構成比率となる評価データを用意し，できるだけ全項目の分類を過不足なく精度評価できるよう考慮した．



分類項目	precision	recall	f1-score	support
00.除外	0.76	0.58	0.66	48
01.目的達成	0.51	0.54	0.52	124
02.不具合	0.64	0.50	0.56	18
03.軽快さ	0.90	0.77	0.83	35
04.信頼性	0.82	0.80	0.81	128
05.操作性	0.73	0.78	0.76	176
06.情報の見やすさ	0.68	0.77	0.72	180
07.表現のわかりやすさ	0.82	0.67	0.74	84
08.継続利用	0.78	0.61	0.68	23
09.役に立つ	0.60	0.67	0.63	139
10.おすすめ	0.69	0.63	0.66	116
11.ないと困る	0.75	0.60	0.67	15
accuracy			0.69	1086
macro avg	0.72	0.66	0.69	1086
weighted avg	0.70	0.69	0.69	1086

図 6 Confusion matrix と各種指標値

モデルの評価には Confusion_matrix を用いて，俯瞰的に分類項目ごとの精度を把握したうえで，accuracy や precision，recall といった機械学習モデルの一般的な評価指標値を確認した（図 6）．

Confusion matrix の縦と横の軸に付与された数字は項目番号を示し，「1」であれば「01.目的達成」というように対応付けている．「縦 0：横 0」，「縦 1：横 1」というように，対角線上の数字が大きな値になる傾向であり，つまりは全体として概ね分類に成功している事がうかがえる．

また，指標値を見ると，項目ごとに分類の得意不得意はあるものの，全体的な accuracy，precision，recall の値はいずれも 0.7 程度と，ランダムに分類分けした場合の正解率 1/12=0.08 と照らせば，まずまずの分類精度を示しているといえる．

次に，文書ベクトルを T-SNE で次元削減し散布図としてプロット．機械学習モデルが各項目を切り分けやすい状況かを可視化した（図 7）．

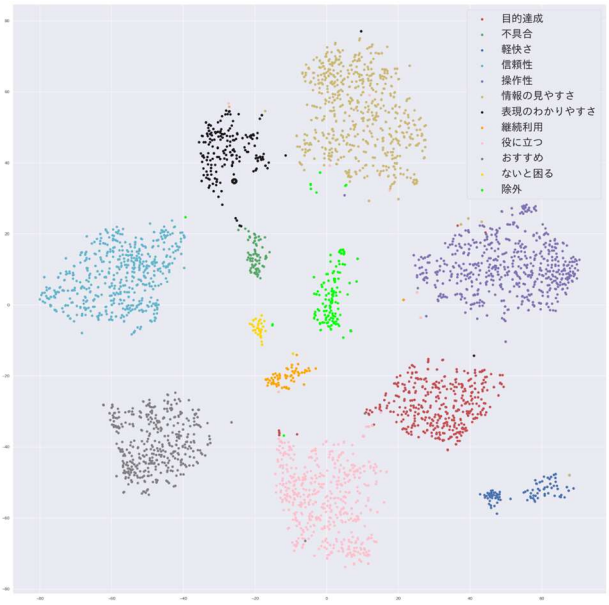


図 7 t-SNE での可視化

図 7 より，どの項目にも該当しない「除外」を含めた全 12 項目が概ね分離できていることから，クラウドソーシングで得られたテキスト群は，機械学習モデルが分類しやすい傾向にあると考えられる．

⑤ 本番データへの適用

今回のモデルは分類精度が 7 割程度であったことから，改めて関係者間で「分類モデルの出力は完璧ではなく，あくまで予測値である」ことを共有．予測が外れているものは担当者が手作業で修正するというハイブリットな運用となるが，分類モデルのアウトプットに分類結果とあわせて予測確率を出力す

ることで，どの分類が誤っている可能性が高く，重点的に目視でチェックしなければならないか，担当者が示唆を得られるように工夫した．

コメント	分類結果	予測確率
ニュースの内容が思想的に偏りがあるように思う	目的達成	0.08270735
見出しがわかりやすい	表現のわかりやすさ	0.97069776
見出しだけでも今日あったことがなんとなくわかる	表現のわかりやすさ	0.65842754
内容がかいつまんであるので短時間でも分かりやすい	情報の見やすさ	0.27513972
慣れているので##特に不満は##ないです。	継続利用	0.09010299
タイトルが適切でないことがよくある##何をいいたいのか分からない記事がある	信頼性	0.4610268

図 8 分類結果のアウトプット形式

4 効果

構築した分類モデルの評価は良好であったため，それを UX 評価の運用へ導入．品質評価部門の担当者の手作業による運用と，分類モデルを用いた運用を並走させ，導入効果を検証することとした．

効果検証は主に以下 2 つの観点で行われた．

検証１：担当者の分類と本モデルの分類がどれだけ一致しているか

検証２：本モデルを導入することでどれだけ作業工数を減らせるか

4.1 効果検証結果

効果検証の結果は表 1 の示す通りとなった．

表 1　モデル導入の効果検証結果

評価指標	評価値
担当者の分類と本モデルの分類の一致率 （1,220 件の本番データを使い一致率を算出）	**71%一致** (867 件一致)
本モデル導入で削減できる工数 （手作業で掛かる工数は 16 時間）	**69%削減** (11 時間削減)

分類精度，削減工数ともに良好であり，さらに定性評価では担当者より長所として以下の点が挙げられた．

- モデルを品質評価部門が自ら運用できそう
- 「予測確率」を使うと分類の補正が捗る
- 「人手で補正したデータ」を使ってモデルを再学習することで精度向上も見込める

なお，一致率 71%の内訳は「付録 A」に掲載する．

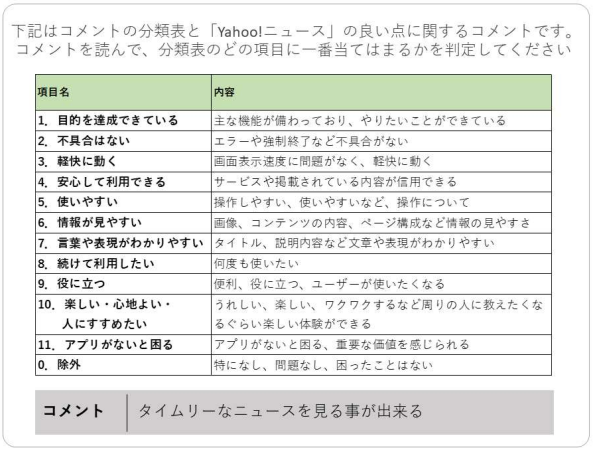
4.2 費用対効果

今回のコメント分類モデルの開発に掛かった工数は約 20 時間である．分類モデルを通じて 1 回あたりの作業工数を 11 時間削減できた事から，単純計算で本モデルを 2 回稼働させれば，開発工数はペイできることになるため，コストパフォーマンスの観点で見ても優れた施策であると言える．

一方で，機械学習モデルを導入せずともクラウドソーシングを使って分類した方が低コストかつ高精度に分類できるのではないかという疑問がわく．

そこで我々は，機械学習モデルで分類したコメントとまったく同じコメントをクラウドソーシングでワーカーに分類を依頼することとした（図 9）．

コメント 1 件につき優良なワーカー10 人に分類させ，多数決で正解を 1 つに決定する方針とした．


下記はコメントの分類表と「Yahoo!ニュース」の良い点に関するコメントです。
コメントを読んで、分類表のどの項目に一番当てはまるかを判定してください

項目名	内容
1. 目的を達成できている	主な機能が備わっており、やりたいことができている
2. 不具合はない	エラーや強制終了など不具合がない
3. 軽快に動く	画面表示速度に問題がなく、軽快に動く
4. 安心して利用できる	サービスや掲載されている内容が信用できる
5. 使いやすい	操作しやすい、使いやすいなど、操作について
6. 情報が見やすい	画像、コンテンツの内容、ページ構成など情報の見やすさ
7. 言葉や表現がわかりやすい	タイトル、説明内容など文章や表現がわかりやすい
8. 続けて利用したい	何度も使いたい
9. 役に立つ	便利、役に立つ、ユーザーが使いたくなる
10. 楽しい・心地よい・人にすすめたい	うれしい、楽しい、ワクワクするなど周りの人に教えたくなるぐらい楽しい体験ができる
11. アプリがないと困る	アプリがないと困る、重要な価値を感じられる
0. 除外	特になし、問題なし、困ったことはない

コメント	タイムリーなニュースを見る事が出来る

図 9 クラウドソーシングのコメント分類タスク

結論から述べると，担当者の分類とクラウドソーシングの分類の一致率は 56%と，機械学習モデルの分類よりも劣る結果となった（表 2）．

表 2　担当者の分類との一致率

一致率を比較する対象	一致率
担当者の分類と機械学習モデルの分類	**71%**
担当者の分類とクラウドソーシングの分類	**56%**

そもそも，この分類作業は UX の知見を有する担当者でも正確な分類は難しく，今回のクラウドソーシングの作業結果を見ても，ワーカーの分類が分散し多数決が成立しないケースが散見された．

このような専門的な観点でのコメント分類はエキスパートの作業を機械学習モデルに学習させるアプローチをとった方が効果を得られやすいと言える．

5 おわりに

fastText は「学習が速い・精度が良い・導入がしやすい」といった特長があるが，今回の取り組みを通じて「関係者との意思疎通も高速化できる」という特長があると実感した．関係者への説明がしやすいことから，学習データ作成を分担する等の協力体制も築きやすく，推論や再学習方法の引継ぎも比較的容易であった．fastText は高精度な自然言語処理を安価かつスピーディーに業務に導入したいというケースに適した手段であると言えよう．

参考文献

1. Yahoo!クラウドソーシング. https://crowdsourcing.yahoo.co.jp/
2. Aarron Walter. Designing for Emotion. 2011.
3. 山﨑 郁未, 伊藤 理紗, 中村 聡史, 小松 孝徳. Web アンケートにおける不真面目回答予防システム実現に向けた自由記述配置の基礎検討, 情報処理学会 研究報告ヒューマンコンピュータインタラクション（HCI）, Vol.2021-HCI-195, No.34, pp.1-8, 2021.
4. fastText. https://fastText.cc/
5. MeCab. Yet Another Part-of-Speech and Morphological Analyzer. https://taku910.github.io/mecab/

付録 A. 効果検証の一致率の内訳

本文「4.1 効果検証結果」で述べた「担当者の分類と本モデルの分類の一致率：71%」の内訳は以下「表 3」の通りであった．全体的に「図 6」と同じ傾向であり，本番データにおいてもモデル開発時と同等の分類精度であった．

表 3 担当者の分類と本モデルの分類の一致率

分類項目	A.手作業で分類した数	B.分類モデルが分類した数	A と B の一致数	一致率 (C/B)
除外	126	72	64	88.9%
目的達成	184	228	133	58.3%
不具合	8	26	7	26.9%
軽快さ	16	15	13	86.7%
信頼性	63	97	57	58.8%
操作性	71	108	62	57.4%
情報の見やすさ	241	240	193	80.4%
表現のわかりやすさ	192	143	129	90.2%
継続利用	3	10	1	10.0%
役に立つ	294	243	188	77.4%
おすすめ	22	37	20	54.1%
ないと困る	0	1	0	0.0%
合計	1220	1220	867	71.1%

付録 B. 安価な施策

付録 B.1 Yahoo!クラウドソーシング

クラウドソーシングとは「crowd=群衆」と「sourcing=委託」からなる造語であり，インターネットを通じて企業が不特定多数の人々に仕事を依頼できるプラットフォームの事を言う．

Yahoo!クラウドソーシングはアンケートや簡易なアノテーションなど誰でも実施可能な「マイクロタスク」に特化したサービスとなっており，ユーザーはタスクを実施することで，PayPay ポイントが獲得できる（図 10）．

図 12 Yahoo!クラウドソーシングの概要

1 サンプル当たりの費用は 10 円ほどであり，例えば 1,000 人に対してアンケートを実施した際に掛かる費用は 1 万円となり「安価なリサーチツール」と言えるだろう．

付録 B.2 機械学習モデルの開発環境

ヤフーには「AI プラットフォーム」と呼ばれる AI・機械学習の開発環境がある．これは，Google Cloud Platform の AI Platform Notebooks のような，マネージド型の JupyterLab 環境であり，ヤフーの社員であれば誰でも簡単にブラウザから python のコーディングと実行が行える．一般的に，AI や機械学習モデルの開発には，サーバーのリソース管理，環境構築，サーバメンテナンス，データアクセスのための準備等に時間と手間がかかるが，AI プラットフォームは利用者がこれらを意識することなく，本来注力するべき AI や機械学習モデルの精度向上やサービス改善業務に注力できる．

GPU の利用も可能であるが，今回は GPU を利用せずに fastText の学習を完了．学習は 1 分足らずで完了したことからも「安価な開発環境」で開発できたソリューションであると言えるだろう．