質問文中の特徴語と感情を用いた半教師あり学習によるトピックの分類

ロ　エイケン† 　井上　潮‡

†東京電機大学 工学研究科 情報通信工学専攻 〒120-8551 東京都足立区千住旭町 5 番
‡東京電機大学 工学部 情報通信工学科 〒120-8551 東京都足立区千住旭町 5 番
E-mail: †21kmc23@ms.dendai.ac.jp, ‡inoueu@mail.dendai.ac.jp

あらまし 多くの Q&A サイトにおいて，ユーザの質問はあらかじめ設定されたカテゴリによって分類されている．しかし，時間の経過に伴い新しいトピックの質問が増加してくると，それに対応した新たなカテゴリを追加することになる．その際に，既存の質問のトピックを調べて，追加したカテゴリに分類し直す必要が生じる．本研究では，質問文中の特徴語と感情を分析することにより，新たなカテゴリに分類する質問を自動的に抽出する手法を提案する．提案手法では，少ない正解ラベル付きデータで学習できる半教師あり学習を用いる．育児・子育ての Q&A サイト「ママリ」に投稿された質問の中から COVID-19 に関する質問を高い精度で抽出できることを確認した．

キーワード 自然言語処理，半教師あり学習，感情分析，テキスト分類

1. はじめに

2020 年に世界中に広がり始めた感染症である新型コロナウイルスは 3 年間続き、さまざまな流行防止戦略が人々の生活様式を大きく変えた。外出自粛や予防接種、在宅勤務などの対策は、社会の一般人だけでなく、高齢者や妊婦、子どもにも影響を与える。未知のウイルスに直面したとき、人々の恐怖や不安を解消する最善の方法の 1 つは、ウイルスに関するさまざまな正しい情報や、ウイルスによって変化した日常生活に関する情報を迅速かつ便利に入手できるようにすることである。

インターネット上には多様な Q&A サイトが開設されている。ユーザーは自分の興味のある分野の Q&A サイトで質問をして、他のユーザーの回答を求めることができる。また、ユーザーは他のユーザーの質問に回答することにより、他のユーザーを助けることもできる。Q&A サイトは大衆の力を活用する相互扶助コミュニティとして、ますます人気が高まっている。新型コロナウイルスが社会に及ぼす影響が大きいことを考えると、Q&A サイトで新型コロナウイルスに関する質問を検索しやすくすることが必要である。多くの Q&A サイトでは検索を支援するために、ユーザーの質問はあらかじめ設定されたカテゴリによって分類されている。しかし、新型コロナウイルスのように新しいトピックに対応したカテゴリを追加しようすると、既存の質問のトピックを調べ直すことが必要になる。

本研究では、妊娠と子育てをテーマにした Q&A サイト Mamari[1]について、既存のカテゴリをベースに、新型コロナウイルスに関する新たなカテゴリを追加することを検討する。これによって、妊娠中の女性と子育ての母親が新型コロナウイルス関連の問題に遭遇したときに、対応する解決策をすばやく見つけることができるようになることが期待できる。なお、本研究で使用したデータは、経営科学系研究部会連合協議会主催の令和 3 年度データ解析コンペティションにおいてコネヒト株式会社から提供された Q&A サイト Mamari のデータのうち、2019 年 1 月から 2021 年 7 月までの質問である。

本稿の構成は以下の通りである。第 2 章では、Q&A サイトと新型コロナウイルスに関する関連研究を紹介する。第 3 章では、特徴語と感情分析を用いて、新しいカテゴリに質問を分類するための手法を提案する。第 4 章では、提案手法の有効性を検証するための実験結果を示し、考察する。第 5 章では、まとめと今後の課題を述べる。

2. 関連研究

Q&A サイトに投稿された質問と回答を分析する研究は活発に行われており、質問の分類手法に関する研究も多い。栗山ら[2]は、投稿された質問を人手で分析することにより、質問をいくつかのタイプに分類し、各タイプの質問を識別するために共通する特徴を抽出した。渡邊ら[3]は、質問に対する客観的な正解や回答の意見性への期待の有無などを考慮した機械学習による質問分類を行った。大森ら[6]は、訓練データに正解情報を人手で付与するコストを削減するために、正解情報を疑似的に付与する訓練データを作成する手法を提案した。加藤ら[4]は、深層学習を用いた質問文のカテゴリ分類手法について分類精度の評価を行い、入力の次元数を増やすことが精度向上に有効であることを確認した。島田ら[5]は、本研究と同じ Q&A サイト Mamari のデータを用いて、共感を求める質問とそうでない質問に対して機械学習を用いて分類を行った。

また、一般のソーシャルメディアに投稿されたテキストの感情分析も盛んであり、最近ではコロナ禍の影響にも注目した研究も発表されている。吉田[7]は、コ

ロナ禍におけるソーシャルメディアの状況について解説している。鳥海ら[8]は、Twitter のデータを用いて新型コロナウイルスに関する大きなイベントの発生とユーザーの投稿にあらわれる感情の関係を明らかにした。峰滝[9]は、２度目の緊急事態宣言発令前後の Twitter 上での反応を時系列で分析し、ネガティブ感情とポジティブ感情の割合差が感染者数の先行指標となることを示した。福田ら[10]は、人々がワクチンに対して持つ安心感や不安感のその要因を解明するために、Twitter 上におけるワクチンに対する人々の感情とその要因を分析した。

機械学習は教師あり学習、教師なし学習、半教師あり学習に大別できる。近年、半教師あり学習に関する研究も多い。江里口ら[11]は、グラフ構造に基づく半教師あり学習の精度向上を目指して、質が高い教師データとグラフ構造を収集する方法を検討した。小暮ら[12]は、商品に関する膨大なカスタマーレビューを分類するために、半教師あり学習とランダムフォレストを組み合わせた分類器の性能を評価した。新納ら[13]は、教師あり学習の教師データとテストデータの領域が異なる問題に対して、半教師あり学習と素性の重み付け学習を交互に適用することにより、分類精度を向上させていく手法を提案した。

3. 提案手法

3.1 データの収集、ラベル付け

本研究では、Mamari のデータのうち covid19 の流行が始まった後の質問、つまり 2020 年 1 月から 2021 年 7 月までの 1 年 7 か月間に投稿された質問を分類対象とする。

最初に、1 年 7 ヶ月間の質問データの一部に手動でラベル付けをした。もともと質問は表 1 のカテゴリに分けられている。そこで各カテゴリの件数の比率に応じて、合計 20,000 件の質問をランダムに抽出し、その中から特徴語である「コロナ」という文字列を含む 1008 件の質問を教師データとした。

表 1: Mamari のカテゴリ id と詳細

id	詳細	id	詳細
1	妊娠・出産	12	お仕事
3	子育て・グッズ	13	ファッション・コスメ
4	サプリ・健康	14	家族・旦那
5	ココロ・悩み	15	お出かけ
6	妊活	16	産婦人科・小児科
7	家事・料理	18	住まい
9	お金・保険	99	その他の疑問
11	雑談・つぶやき		

そして、質問を新型コロナウイルスに関連するものと関連しないものの 2 つに分けた。ここでコロナに関連しているかどうかを判断する基準は、コロナの悪影響を受けているかどうかとした。例えば「夫は在宅勤務なので長く付き合えるから、よりラブラブ」、「人との交流が嫌いなので、コロナのせいで一人でいると気分が良くなる」などの質問はコロナに関連しないものとした。

次に、コロナに関連する質問は、もともとのカテゴリに基づいて、id:1（妊娠・出産）、4（サプリ・健康）、16（産婦人科・小児科）の質問を健康疾患に関連するカテゴリ、id:5（ココロ・悩み）、14（家族・旦那）の質問をメンタルヘルスに関するカテゴリ、残りのカテゴリの質問を「その他」のカテゴリに分けた。つまり、1008 件の質問には 4 つのラベルのいずれかが付与されたことになる。各ラベルの件数と意味を表 2 に示す。

表 2: ラベルの件数と意味

ラベル	件数	意味
0	102	コロナと関係ない
1	154	健康・病気
2	214	メンタル・悩み
3	538	その他の疑問

3.2 感情分析スコアの差

3.2.1 「コロナ」を含む文の削除

自然言語処理の分野では、データ分析のために感情分析の手法がよく使用される。たとえば、テキストがポジティブ感情かネガティブ感情かを判断するのは最も基本的なものである。本研究では、感情辞書（感情を既にマークした辞書）を使用して個々の単語に感情のスコアを付け、質問中のすべての単語のスコアを加算することにより質問の感情スコアを求めた。ただし、「コロナ」による悪影響を受けているかどうかを判断するため、本文中に「コロナ」を含む文を削除した時の感情スコアと、元のテキストの感情スコアの差を計算した。ここで重要な点は、文の終わりの判定方法である。文の終わりに句点「。」以外の記号が使われることがあるため、読点「、」以外の記号を「。」に置き換えた。具体例を図 1 に示す。

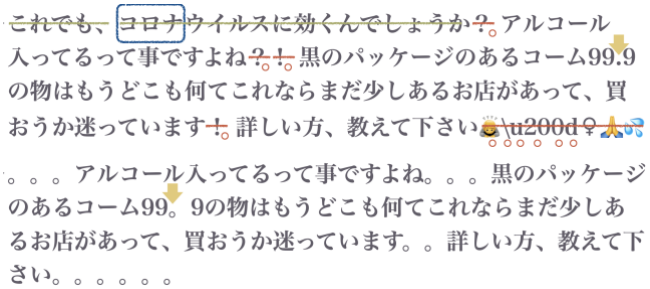

図 1 :「コロナ」を含む文の削除方法

3.2.2 感情分析スコアの差の計算

本研究における感情分析スコアの差の計算は、BERT の学習済み日本語の感情分析モデル（bert-base-

japanese-sentiment）[14]を使用した。

BERT モデルでテキストデータを処理する場合、最大長は 512 バイトである。しかし、Mamari では、ユーザーの入力長に制限がないため、512 バイトを超える質問が多くある。512 バイトを超える質問については、元の質問文の前後各 256 バイトを取得して新しいテキストにマージする方法を行う。

BERT の感情分析モデルはテキストをポジティブかネガティブの 2 つの感情に分け、0.5 から 1 までの感情スコアを付けたものである。例えば、「私は幸福である。」という文の感情スコアを図 2 に示す。

```
("私は幸福である。"))
```

```
[{'label': 'ポジティブ', 'score': 0.9934489130973816}]
```

図 2：BERT 感情分析モデル出力例

感情のラベルはポジティブとネガティブの 2 種類があり、「コロナ」を含む文を削除する前後での組み合わせによって、図 3 に示す計算方法で感情分析スコアを求めた。

ポジ→ポジ：nocovid_score - covid_score

ポジ→ネガ：(1 - nocovid_score) - covid_score

ネガ→ポジ：nocovid_score - (1 - covid_score)

ネガ→ネガ：covid_score - nocovid_score

nocovid_score :「コロナ」を含む文を削除した感情分析のスコア

covid_score :「コロナ」を含む文を削除しない感情分析のスコア

図 3: 感情分析スコアの計算方法

3.2.3 感情分析スコアの差の分析

最終的に算出された感情分析スコアの差の範囲は-1～1 で、1 に近いほどポジティブ、-1 に近いほどネガティブとなる。1008 件の質問の感情分析スコアの変化を表 3 に示す。

表 3: 感情分析スコアの変化

感情変化	件数
posi→posi	760
posi→nega	30
nega→nega	86
nega→posi	132

感情のラベルが変化しない質問を含めて、1008 件の質問のうち 836 件の感情分析スコアの差が正数で、全体の 82.9% を占めている。したがって、特徴語「コロナ」は悪影響があると推測できると考えられる。

3.3 半教師あり学習

半教師あり学習は、少数の教師データを利用して多数のデータに疑似的なラベル付けを行うことによって機械学習の精度向上を図るものである。本研究では、scikit-learn のラベル拡散法 (label spreading) [15]を使用した。

教師データは 3.1 と 3.2 で述べた人手によるラベルと感情スコアの差が付与された 1008 件の質問データである。テストデータは、2020 年 1 月から 2020 年 12 月までの期間でランダムに選択された 2 千件、5 千件、1 万件と 2 万件の「コロナ」というキーワードを含まない質問データである。抽出したテストデータの件数は、教師データのそれぞれ 2 倍、5 倍、10 倍と 20 倍に相当する。そして、テストデータのラベルの初期値は −1 に設定し、半教師あり学習によって表 2 に示した 0 から 3 までのいずれかのラベル値に更新する。

4. 評価

4.1 基本となるアルゴリズム

本研究では、比較のために 6 つの分類アルゴリズムを使用した。それらは k-NN（k 近傍法）、DT（決定木）、MLP（多層パーセプトロン）、SVM（サポートベクターマシン）、RF(ランダムフォレスト)、GBT（勾配ブースティング）である。アルゴリズムの厳密な性能比較ではないため、各アルゴリズムのパラメータの特別なチューニングはせず、最も基本的な値を使用した。各アルゴリズムの主要なパラメータを表 4 に示す。

表 4: 各アルゴリズムのパラメータ

アルゴリズム	主要なパラメータ
k-NN	k=3
DT	max_depth = 10
MLP	hidden_layer_sizes = (100), activation = 'relu'
SVM	kernel = ‘poly’, degree = 3
RF	max_depth = 10, n_estimators=100, random_state=2
GBT	max_depth = 3

4.2 アルゴリズムの比較

手動でラベル付けをした教師データ 1008 件を使用して、感情分析スコアを用いない条件での各アルゴリズムの分類精度を評価した。汎化性能を評価するため、10 分割交差検証を行った。各アルゴリズムのテストデータの平均分類精度(accuracy)を表 5 に示す。

表 5: 各アルゴリズムの分類精度

アルゴリズム	分類精度
k-NN	0.442
DT	0.690
MLP	0.616
SVM	0.715
RF	0.616
GBT	0.678

表 5 において、SVM の分類結果が最良であり、DT と GBT の分類結果はやや劣るが比較的良好で、RF と

MLP の分類結果はやや悪く、k-NN の分類結果は最悪である。

4.3 感情分析スコアの追加

感情分析スコアの差を追加した結果を表 6 と図 4 に示す。各アルゴリズムの分類精度を追加前と比較すると、k-NN、DT、MLP と RF の 4 つのアルゴリズムの精度が上がり、SVM と GBT の 2 つのアルゴリズムの精度が下がった。その中で、k-NN の精度向上が最も大きく、約 0.05 であり、他の 3 つのアルゴリズムの向上は約 0.01 である。分類精度が低下した 2 つのアルゴリズムは、それぞれ 0.005 と 0.002 でわずかであった。

表 6: 感情分析スコア追加した分類精度の対比

アルゴリズム	感情分析なし	感情分析あり
k-NN	0.442	**0.494**
DT	0.690	**0.699**
MLP	0.616	**0.625**
SVM	**0.715**	0.710
RF	0.616	**0.630**
GBT	**0.678**	0.676

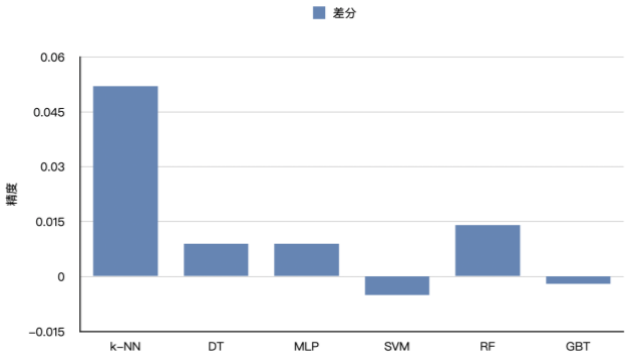

図 4: 感情分析スコア追加した分類精度の差分

以上の結果から、感情分析スコアを追加することにより、多くの場合に分類精度が向上することが分かった。しかし、分類精度のわずかな低下が発生する場合もある。

4.4 ラベルなしデータの分類

Mamari に投稿された質問のうち、手動でラベル付けしたものは全体のごく一部に過ぎない。このような少数のデータで学習したモデルで、大量のデータを分類できるかどうかを検証する必要がある。そこで、2020 年 1 月から 2020 年 12 月までの期間で特徴語である「コロナ」を含む 10,000 件の質問をランダムに選択してテストデータを作り、前節と同様の評価を行った。ただし、データはラベル付けがされていないため、分類結果の中から 100 件の質問をランダムにサンプリングし、分類精度を手動で評価した。評価結果を表 7 に示す。

表 7: ラベルなしデータの分類精度

アルゴリズム	感情分析なし	感情分析あり
k-NN	0.60	**0.62**
DT	0.65	**0.69**
MLP	**0.76**	**0.76**
SVM	**0.71**	0.70
RF	0.70	**0.76**
GBT	0.75	**0.76**

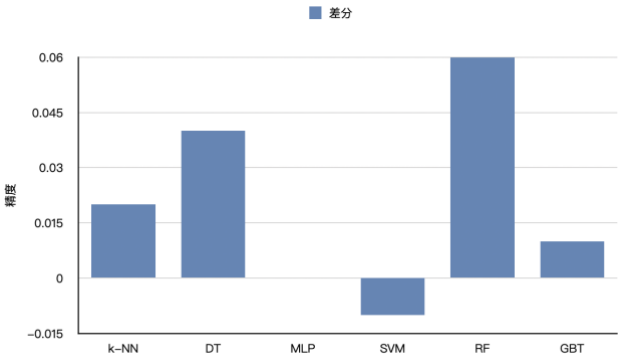

図 5: ラベルなしデータの分類精度の差分

表 7 では、GBT と MLP が最も良い分類精度となった。トレーニングデータの数が少ないにもかかわらず、比較的良好な分類精度が得られた。全体として、感情スコアの追加により、分類精度が改善されたことが分かった。

なお、カテゴリによって分類精度に違いがでる。たとえば、いくつかのアルゴリズムは、表 2 のラベル 2 の質問をより正確に分類できた。将来的には、異なるアルゴリズムの結果を組み合わせて、より高い予測結果を持つモデルを作成できる可能性がある。

4.5 半教師あり学習の結果

4.4 までの実験では、6 種類のテキスト分類アルゴリズムを評価した。ラベル付きと感情スコアのデータを教師データとして、一万件のラベルなしのテストデータを分類した結果、精度向上の大きさは k-NN、RF の順であった。しかし、k-NN は精度向上後の値も他よりも低いため、半教師あり学習は RF を使用した（パラメータは同じ）。

半教師あり学習の分類精度を評価した結果を表 7 に示す。テストデータの件数が教師データの 10 倍のときに最も良い分類精度となり、表 7 の最高の分類精度を上回った。しかし、20 倍になると分類精度が下がった。

表 7:半教師あり学習の分類精度

倍数	分類精度
2	0.779
5	0.806
10	0.824
20	0.810

テストデータが教師データの 10 倍の条件での分類精度の変化を図 6 に示す。半教師あり学習によって分類精度が向上することが確認できた。

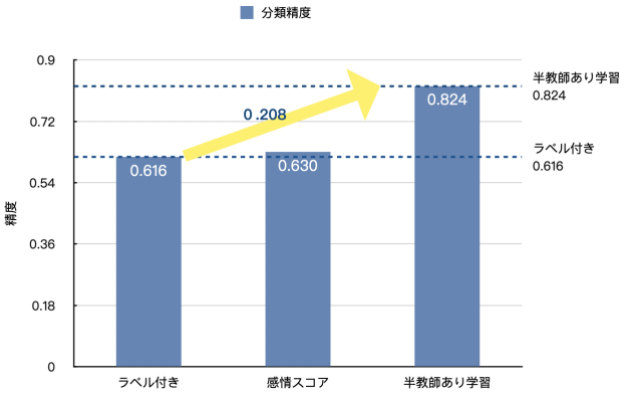

図 6: RF の分類精度の変化

5. まとめ

本研究では、Q&A サイトにおける質問を分類するための感情分析を用いた新しい手法を提案した。従来の感情分析との違いは、新たなカテゴリを表していると考えられる「コロナ」などの新語を特徴語として捉え、特徴語を含む文を削除した時に生じる感情スコアの変化を計算する点にある。Mamari に投稿された質問の一部に人手でラベル付けを行い、一般的に使用されている 6 つの分類アルゴリズムについて分類精度を評価した結果、提案手法による分類精度の向上を確認できた。

また、ラベル付けされていないデータの分類精度を向上させるため、RF(ランダムフォレスト)を使用する半教師あり学習を導入することにより、分類精度がさらに向上することを確認できた。

今後は、「コロナ」以外の特徴語に対しても有効性を評価する必要がある。

謝　辞

本研究で扱ったデータは、経営科学系研究部会連合協議会主催の令和 3 年度データ解析コンペティションで提供されたものである。関係各位に深く感謝する。

参　考　文　献

[1] ママリ｜ママの一歩を支える情報サイト, https://mamari.jp
[2] 栗山和子, 神門典子, Q&A サイトにおける質問と回答の分析, 情報処理学会研究報告, Vol.2009-DBS-148, No.19, pp.1-8, 2009.
[3] 渡邊直人, 島田諭, 関洋平, 神門典子, QA コミュニティにおける質問者の期待に基づく質問分類に関する一検討, DEIM2011, B5-1, 2011.
[4] 加藤玲大, 馬青, 村田真樹, 深層学習を用いた QA サイト質問文のカテゴリ分類, 情報処理学会研究報告, Vol.2016-NL-228, No.10, pp.1-6, 2016.
[5] 島田達朗, 櫻井彰人, コミュニティサイトにおける共感を求める質問の認識, 知能と情報, Vol.29, No.4, pp.611-618, 2017.
[6] 大森勇輔, 森田和宏, 泓田正雄, 青江順一, 疑似訓練データを用いた Q&A サイトの質問分類, 言語処理学会 21 回年次大会 発表論文集, pp.489-492, 2015.
[7] 吉田光男, COVID-19 流行下におけるソーシャルメディア−日本での状況と研究動向・公開データセット−, 人工知能, Vol.35, No.5, pp.644-653, 2020.
[8] 鳥海不二夫, 榊剛史, 吉田光男, ソーシャルメディアを用いた新型コロナ禍における感情変化の分析, 人工知能学会論文誌, Vol.35, No.4, F, pp.1-7, 2020.
[9] 峰滝和典, Twitter 上の新型コロナウイルス関連語句の分析―2 度目の緊急事態宣言前後の動向に焦点をあてて―, 商経学叢 , Vol.67, No.3, pp.101-120, 2021.
[10] 福田悟志, 難波英嗣, 庄司裕子, コロナ禍におけるワクチンに対する人々の感情変化とその要因の分析, 情報処理学会研究報告, Vol.2022-IFAT-145, No.2, pp.1-6, 2022.
[11] 江里口瑛子, 小林一朗, 半教師あり学習における教師データ選出とグラフ構成, 人工知能学会 インタラクティブ 情報アクセスと可視化マイニング研究会（第 3 回）, SIG-AM-03-05, 2013.
[12] 小暮枝里子, 齊藤史哲, 石津昌平, ランダムフォレストの半教師あり学習による“顧客の声”の分類, 日本感性工学会論文誌, Vol.17, No.5, pp.537-545, 2018.
[13] 新納浩幸, 古宮嘉那子, 佐々木稔, 半教師あり学習と素性の重み付け学習の交互適用による文書分類の領域適応, 言語処理学会, 第 22 回年次大会発表論文集, 2016.
[14] BERT 学習済み日本語の感情分析モデル, https://huggingface.co/daigo/bert-base-japanese-sentiment, 2022 年 10 月 5 日閲覧
[15] scikit-learn のラベル拡散法 (label spreading), https://scikit-learn.org/stable/modules/generated/sklearn.semi_supervised.LabelSpreading.html, 2023 年 2 月 15 日閲覧