

Twitterの複数アカウント所持者を対象とした 投稿アカウント推定手法の提案

野崎 祐里[†] 佐藤 哲司^{††}

[†] 筑波大学 情報学群 知識情報・図書館学類 〒305-8550 茨城県つくば市春日 1-2

^{††} 筑波大学 図書館情報メディア系 〒305-8550 茨城県つくば市春日 1-2

E-mail: †{nozaki,satoh}@ce.slis.tsukuba.ac.jp

あらまし Twitterはプロフィールを実名で登録する義務がないため、用途に応じて複数のアカウントを使い分けるユーザが存在する。このような複数アカウント所持者は、投稿するアカウントを間違えることでトラブルを引き起こすことがある。本研究では、機械学習を用いてツイートの投稿アカウントを推定する手法を提案する。機械学習の手法はSVMとRandom Forestのいずれかとし、学習に使う素性は投稿内容や投稿形態を反映した素性と著者推定で利用されている素性を用いる。機械学習の手法と素性を組み合わせた評価実験の結果、投稿内容や投稿形態を反映した素性をSVMで分類した際に最も高い精度で分類することができたので報告する。

キーワード Twitter, 機械学習, マルチアカウント

1. はじめに

Twitter^(注1)はツイートと呼ばれる短文を投稿するマイクロブログである。Twitterの特徴として、実名で登録する必要はなくユーザの好きな名前でも登録することができる。そのため、裏アカウントや趣味用など複数のアカウントを作成し、目的に応じて異なるアカウントを使い分けることが可能である。総務省の平成27年度版情報通信白書[1]によると、SNSの1つであるFacebook^(注2)において、実名利用と回答したユーザは84.8%であったのに対し、Twitterは23.5%であった。また、モバイルマーケティングデータ研究所が2011年に行った登録アカウント数の調査[2]では、Facebook利用者の8.9%が複数アカウントを所持している一方で、Twitterは25.8%と約4人に1人の割合で複数アカウントを利用していることを明らかにしている。私的なアカウントを複数所持していなくても、自分が所属している企業やサークルのアカウントを担当し、私的なアカウントと公式アカウントを管理することがある。

このような複数のアカウントを利用しているユーザは、投稿するアカウントを間違えてツイートしてしまうことが考えられる。アカウントを間違えて投稿してしまうと、実名で利用しているアカウントとそうでないアカウントとの関連付けが推定されることや、公式アカウントに私的な内容をつぶやいて信用を失うことがある。2016年に開催された全国高校野球選手権大会で、試合中の判定を巡り、新聞社のツイッター公式アカウントに、監督を中傷する文章が同社記者から投稿され、新聞社は「私的なやりとりが誤って投稿された」として謝罪した事件があった[3]。一度誤ったツイートを投稿をしてしまうと、多くのユーザがそれぞれに広めてしまい、ツイートを削除しても手遅

れになっていることがある。

投稿アカウントの間違いを防止する方法として、小林[4]はアカウントごとに使用する投稿ツールや投稿端末を別にすることを提案している。しかし、この方法では気軽に投稿できるというTwitterの特性を潰してしまうことが考えられる。そこで本稿では機械学習の手法により、適切な投稿アカウントを推定する手法を提案する。複数アカウント所持者は特定の目的に応じてアカウントを使い分けていることに着目し、投稿内容や投稿形態を反映したものを素性として抽出する。そして、複数アカウント所持者のツイートを収集して評価実験を行うことにより、提案手法の妥当性を明らかにする。また、分類正解率だけではなく、アカウントの使い分け方によって機械学習が予測する結果の方向性に差があるのかや、分類に有効な素性の種類についても検証する。

本論文の構成は次の通りである。第2章では、関連研究を概観し、本研究の位置づけを示す。第3章では提案手法として、複数アカウント所持者の投稿分類をするために、機械学習で利用する素性について説明する。第4章で、複数アカウント所持者の投稿データを対象とした分類実験を行い、提案手法の有用性を評価する。第5章で、分類実験で算出したそれぞれの結果について考察する。最後に第6章で、本論文のまとめと今後の課題を述べる。

2. 関連研究

本研究は、Twitterの複数アカウント所持者のツイートを機械学習の手法で分類するものである。関連研究として、SNSの使い分けや、文書分類の研究が挙げられる。文書分類に関しては、1つのアカウントを1人の著者と考え、ツイートがどの著者が投稿したものを推定する著者推定の研究に近い。また、複数アカウント所持者は特定の目的に応じてアカウントの使い分けをしていると考えられる。目的によって投稿される文書の

(注1) : <https://twitter.com/>

(注2) : <https://www.facebook.com/>

内容が変わることから、話題推定の研究に近いともいえる。

澤田ら [5] は Twitter の利用者に対して、アカウントを複数持っているか、複数所持している場合の使い分けについて質的調査を行っている。調査の結果、Twitter の利用者 20 人中 12 人が複数アカウントを所持していた。また、アカウントの使い分け方として、趣味や所属するコミュニティに応じた使い分けや、愚痴用や裏アカ^(注3)としての使い分けのほか、日記用や情報収集専用アカウントなどを設けていることが明らかにしている。木原ら [6] は、Twitter とブログという異なるソーシャルメディアのアカウント関連付け防止システムの構築を行っている。具体的には、ユーザが新しくブログ記事を投稿する際に、ユーザの所持している Twitter アカウントとの類似度が高い場合に警告を与え、表現をマスキングする仕組みを導入している。類似度を算出するための素性は、形態素の出現頻度、内容語と記号それぞれの tf-idf 値、ユーザ特有の素性、投稿時間である。

Schwartz ら [7] は、英文を対象としたツイート著者推定を行っている。素性は文字 n-gram と単語 n-gram に加え、高頻出の単語の並びの間に低頻出の単語が入る組み合わせが一定以上出現したものを著者特有の表現としている。これらの素性を SVM にかけることで高い精度で著者推定が行えるとしている。奥野ら [8] は、文字 n-gram を素性とした Twitter 著者推定の手法を提案している。n の値は 1~3 で、n の値が大きい素性ほど高い重みづけをしている。また、教師なし学習の手法である k-means を用いて話題にクラスタリングし、分類対象データと最も類似している話題の学習データを選択することによって、分類精度の向上を図っている。著者の判定には機械学習ではなく、文書間の類似度スコアを用いている。

六瀬ら [9] は、東日本大震災のツイート群を、機械学習で津波、避難、ライフライン、交通機関、道路のカテゴリに分類している。素性は名詞のみとし、機械学習の手法は SVM, Random Forest, Naive Bayes 分類器の 3 手法で比較した結果、どのカテゴリも SVM が最も高い精度であった。また、形態素解析器ではうまく抽出できない鉄道路線名や駅名を Web 上から入手し、これらの素性が出現するかどうかを表す 2 値の素性を追加することによって分類性能の向上が見られた。Sriram [10] らは、Naive Bayes 分類器でツイートをニュース、イベント、意見、取引、私的なメッセージの 5 つに分類している。分類の素性には、略語、時間を表す語、意見を表す語、強調を表す語、お金や割合を表す語、先頭の@、ツイート中の@、ツイートの著者の 8 種類を利用している。この 8 種類の素性を用いた分類正解率は、単語を素性とした場合よりも高い結果となった。

同一著者の複数アカウントを対象とした投稿分類に関する研究はまだ知られていないため、本研究は新規性のある研究であるといえる。そのため、基本的な分類手法を実施して基準となる正解率を出し、その基準を超えるような手法を提案する。

3. 提案手法

本研究は Twitter の複数アカウント所持者の投稿するアカウントを機械学習によって推定する。機械学習を利用するためには、素性を抽出してベクトルで表現する必要がある。ベクトル化は、Bag of Words(BoW) で行う。これは、各素性の出現順序を考慮せず、出現頻度をベクトルの形式で表したモデルである。

素性には、ツイート中に出現した形態素を用いる。形態素の種類は、名詞、動詞、形容詞、副詞、記号とする。形態素の抽出には日本語形態素解析器の Mecab [11] を使用する。品詞の中で活用のある動詞と形容詞は、原形に戻して抽出する。また、数字のみの文字列、1 文字だけのひらがな、カタカナ、アルファベットはストップワードとして除外する。

形態素解析器で素性を抽出する手法には 2 つの問題点がある。1 つ目は、1 つのツイートを 1 文書とみなして分類を行うため、特定の品詞の形態素のみを素性とした場合に手がかりとなる素性の数が少なくなってしまうことである。2 つ目は、Twitter のようなくだけた表現が多く出現する文章では、形態素解析器で正確に形態素を抽出することができないことである。そこで形態素に加えて、形態素解析器以外の方法で取り出した素性を追加することによって分類正解率の向上を目指す。複数アカウント所持者は目的に応じてアカウントの使い分けをするが、アカウントの使い分け方によって投稿形態も変化すると考える。投稿形態とは、ツイート中の文字数、リプライ、リツイート、ハッシュタグ、URL や画像の有無のことである。例えば、趣味用のアカウントとして猫について専用につぶやくアカウントがあった場合、そのアカウントの投稿形態として猫に関する画像や URL、ハッシュタグが付与されやすいといえる。さらに、形態素以外で文書の内容を表現する素性として、ツイートの単語感情極性のスコア、ハッシュタグのキーワード、アカウントのスクリーンネーム、URL のドメインも追加する。表 1 に形態素解析器で抽出できる素性(基本素性)、表 2 に拡張素性の種類の一覧を提示する。拡張素性は、Twitter REST API [12] と正規表現を利用して抽出する。ツイートの文字数は、Twitter で入力できる最大の文字数 140 で割ることで正規化を行った。単語感情極性は、高村ら [13] が作成した単語感情極性対応表を利用する。これは、名詞、動詞、形容詞、副詞の単語を $-1 \sim +1$ の実数値で割り当てたもので、 -1 に近づくほどネガティブな単語、 $+1$ に近づくほどポジティブな単語になっている。ツイートの単語感情極性スコア s は式 (1) で定義する。

表 1 基本素性の種類

基本素性
名詞
動詞
形容詞
副詞
記号

(注3)：裏アカとは「裏アカウント」の略称で、本来のアカウント(表アカウント)では言えないような内容を投稿するために設けられたアカウントを指す。

表 2 拡張素性の種類

拡張素性
ツイートの文字数
リプライの有無
リツイートの有無
ハッシュタグの有無
URLの有無
画像の有無
単語感情極性のスコア
アカウントのスクリーンネーム
URLのドメイン
ハッシュタグのキーワード

$$s = \frac{\sum_{i=1}^n w_i}{n} \quad (1)$$

w_i はツイート中の単語 i の単語感情極性スコアで、 n はツイート中に出現した単語感情極性対応表の単語数である。

4. 評価実験

4.1 実験概要

提案手法の有効性を確かめるため、Twitter の複数アカウント所持者の過去のツイートを分類する評価実験を行う。まず、実験用のデータセットとして Twitter 複数アカウント所持者のツイートを用意する。本研究では、プロフィール文に別のアカウントのスクリーンネームが記載されているアカウントと、記載された先のアカウントのペアを、同一著者の複数所持アカウント (マルチアカウント) とする。その際、双方のアカウントが同一著者のものであるか、bot ではないかを人手で確認する。本研究では 2 種類の使い分けのみを対象とし、アカウントを 3 つ以上使い分けしているユーザが存在した場合は、その中の 2 ペアを選択する。

次に、条件を満たしたマルチアカウントを使い分け方に応じてラベル付けを行う。本研究では、使い分け方として公式-普段、裏アカ-普段、趣味-普段の 3 種類に設定した。ここで定義する「公式」は、お店のアカウントや業務用のアカウントのことである。「裏アカ」とは、普段は言えないことや愚痴をつぶやくアカウントのことで、「趣味」は、特定の分野に特化した情報の収集や発信を目的にしたアカウントのことである。ラベル付けは、アカウント名やプロフィール文に記述されている情報をもとに人手で判断する。

ラベル付けに判断したプロフィール文の表現の例

(公式) 公式アカウント, 店名や団体名のアカウント
 (裏アカ) 裏アカ, 裏垢, 裏アカウント, 愚痴用
 (趣味) ○○用のアカウント, 趣味用, スポーツ用
 (普段) 日常用, 普段用, 中の人, 管理人

実験用のデータセットとして、2016 年 8 月 13 日 19 時時点で条件を満たしたマルチアカウントを、公式-普段、裏アカ-普段、趣味-普段の 3 種類の使い分け方でそれぞれ 20 ペアずつ、

計 60 ペアを用意した。そして、それぞれのアカウントから投稿された時間が新しい順に 1000 件ずつツイートを収集した。収集したツイート群から第 3 章の手法で素性を抽出し、それぞれのマルチアカウントごとに分類器を構築する。機械学習は、Support Vector Machine(SVM) と Random Forest(RF) の 2 種類の手法を利用し、それぞれの結果を比較する。

SVM は、2 クラス間のデータとの距離 (マージン) が最大となるような超平面を構築する。SVM の実装にはオープンソースの機械学習のライブラリである LIBSVM [14] を利用した。カーネルは線形カーネルを選択し、コストパラメータ C の値は 1.0 とした。

RF は複数の決定木を構築し、それぞれの弱識別器が判別した結果の多数決によって分類を行うアンサンブル学習のアルゴリズムである。RF では各素性の重要度を、分岐に使われた際の不純度の減少量によって評価することができる。実装には Python の機械学習ライブラリ scikit-learn [15] を利用した。パラメータである決定木を構築するための素性の数は、素性の次元数の平方根の値を、分岐の基準はジニ係数を使用する。決定木は 2 分木を用いる。

4.2 評価方法

評価指標として、5 分割交差検定を行ったときの分類正解率を算出する。分類正解率は、分類器がどれだけうまく分類できるかを表す指標である。分類正解率は式 (2) で表せる。

$$\text{分類正解率} = \frac{\text{分類に正解した評価事例数}}{\text{評価事例数}} \quad (2)$$

それぞれのマルチアカウント (計 60 ペア) の分類正解率を平均した値を全体の結果として提示する。また、ラベル付けした公式-普段、裏アカ-普段、趣味-普段の 3 種類の使い分けごとにも、平均分類正解率を算出する。さらに、5 分割交差検定で、1 アカウントにおけるツイート件数を (50, 100, 200, 400, 600, 800, 1,000) と変化させたときの平均分類正解率の変化も評価する。

評価実験に用いる素性は、基本素性のみの場合、基本素性と拡張素性を組み合わせた場合 (提案素性) に加え、文字 n -gram と比較を行う。文字 n -gram とは隣り合って出現した n 文字の文字列のまとまりのことである。文字 n -gram は文末表現や言い回し、つづりの間違いを考慮するため、著者の癖が出現しやすいと考えられており、Schwartz ら [7] や奥野ら [8] の研究のように、著者推定の手法で多く用いられている素性である。Schwartz ら [7] は英語の文章に対して n の値を 4 としているが、本研究は日本語の文章を対象にしているため、文字 n -gram の n 値は松浦ら [16] の研究を参考に $n=3$ に設定する。また、それぞれの手法で出現した素性の次元数も検証する。

分類に失敗したときのリスクはアカウントの使い分けによって異なることが考えられる。例えば、普段用の内容を裏アカに間違えるよりも、裏アカの愚痴や文句が含まれたツイートを普段用のアカウントに間違える方がリスクが大きいと予想される。そこで、提案素性において、誤りの方向性を表す分類誤り率を算出する。分類誤り率は式 (3) で表せる。

$$\text{分類誤り率} = \frac{\text{誤った方向性の事例数}}{\text{分類に失敗した評価事例数}} \quad (3)$$

分類誤り率は、3種類の使い分けごとに集計し、平均した値を評価する。

そして、提案素性で1アカウントごとのツイート数が1,000のとき、RFを実施した際の素性の重要度上位5件を分類器ごとに取得し、その素性の種類を使い分けごとに集計する。素性の重要度はジニ係数の減少度の大きさで表せる。ジニ係数は、以下の式(4)で定義される。

$$I(t) = \sum_{i=1}^K P(C_i|t)(1 - P(C_i|t)) = 1 - \sum_{i=1}^K P(C_i|t)^2 \quad (4)$$

$P(C_i|t)$ は、ノード t で i 番目のクラスに分類される確率である。ジニ係数の減少度は式(5)で定義される。

$$\Delta I(t) = I(t) - (p_L I(t_L) + p_R I(t_R)) \quad (5)$$

p_L は2分木で分岐された後の左側のノードの分岐割合、 p_R は2分木で分岐された後の右側のノードの分岐割合である。 $\Delta I(t)$ の値が大きいノードに用いられている素性が重要度の高い素性になる。

4.3 実験結果

まず、各手法における素性の次元数を表3に示す。これは、各マルチアカウントに出現した素性の次元数の平均値である。文字3-gramの素性の次元数は、ほかの2つの手法よりかなり多くなっている。素性の拡張により、基本素性と比べて次元数が平均約600程度増加した。

4.3.1 分類正解率

1アカウントにおけるツイート件数を1,000としたとき、5分割交差検定で算出した個々のマルチアカウントの分類正解率を平均した値(全体)と、3種類の使い分けごとに集計して平均した値を表4~7に示す。全体の平均分類正解率は、SVMでは基本素性と拡張素性を組み合わせた提案素性が3種類の素性の中で最も高い正解率になり、基本素性と文字3-gramの正解率は同じになった。RFでは提案素性、基本素性、文字3-gramの順に平均分類正解率が高かった。全体の平均分類正解率で最も高い値になったのは、提案素性をSVMで分類したときで、その値は0.828である。

使い分けごとに平均した分類正解率では、公式-普段が最も高い正解率を出し、その値は約9割であった。裏アカ-普段と趣味-普段が8割弱の正解率で、RFでは裏アカ-普段の正解率などの素性においても趣味-普段より高い結果となった。拡張素性

表3 各手法における素性の次元数の平均値

基本素性	提案素性	3-gram
6,990	7,591	50,936

表4 平均分類正解率(全体)

	SVM	RF
基本素性	0.823	0.809
提案素性	0.828	0.815
3-gram	0.823	0.806

表5 平均分類正解率(公式-普段)

	SVM	RF
基本素性	0.894	0.882
提案素性	0.903	0.889
3-gram	0.889	0.885

表6 平均分類正解率(裏アカ-普段) 表7 平均分類正解率(趣味-普段)

	SVM	RF
基本素性	0.781	0.774
提案素性	0.787	0.783
3-gram	0.787	0.775

	SVM	RF
基本素性	0.795	0.771
提案素性	0.796	0.774
3-gram	0.784	0.76

を加えたことによる正解率の向上で一番大きかったのは、SVMでは公式-普段、RFでは裏アカ-普段でその値はどちらも0.009であった。また、基本素性と文字3-gramの素性を比較すると、公式-普段や裏アカ-普段ではわずかに文字3-gramの方が高いが、趣味-普段では基本素性の方が文字3-gramより0.01以上高くなっている。機械学習の手法による平均分類正解率の違いは、全ての項目に対してSVMがRFを上回る結果になった。

1アカウントにおけるツイート件数を変えたときの平均分類正解率は図1~8の通りである。提案素性では、ツイート数が50でもSVMでは全体の正解率が0.775、RFでは0.754の正解率で分類できた。SVM、RFどちらにおいてもツイート数が増えるにつれて分類正解率が高くなる傾向があるが、ツイート数がある程度まで増えるとほとんど正解率が上昇しないことがある。また、RFの公式-普段の提案素性のツイート数が400から600になるときや、RFの裏アカ-普段のツイート数が800から1,000になるときなど、一部分類正解率が下がる箇所が存在する。全体的にRFの方がツイート数による分類正解率の上昇が大きくなっている。提案素性の分類正解率を上回った素性は、SVMの裏アカ-普段でツイート数が50、200、400のときの文字3-gramと、SVMの普段-趣味でツイート数が200のときの基本素性だけである。RFにおいては全てのツイート数において提案素性が他の素性と同等あるいは上回る結果となった。

提案素性と基本素性の分類正解率の差に注目してみると、ツイート数が50のときのRFの裏アカ-普段の使い分けが0.029と最も高い正解率の向上が見られた。学習データが少ないときのRFにおいて、拡張素性が有効に働いているといえる。

4.3.2 分類誤り率

提案素性における分類誤り率の平均値(平均分類誤り率)を

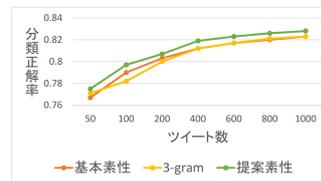


図1 全体(SVM)

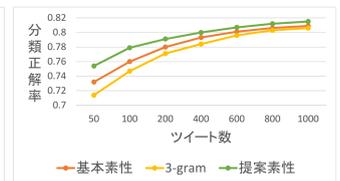


図2 全体(RF)

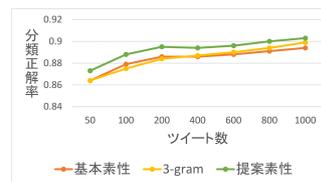


図3 公式-普段(SVM)

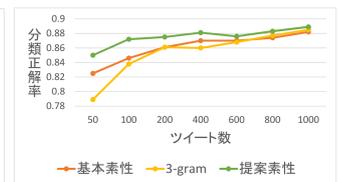


図4 公式-普段(RF)

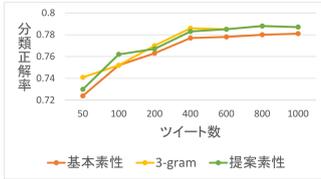


図 5 裏アカ-普段 (SVM)

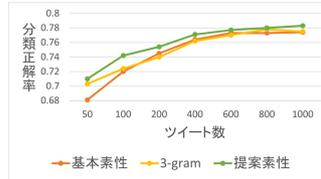


図 6 裏アカ-普段 (RF)

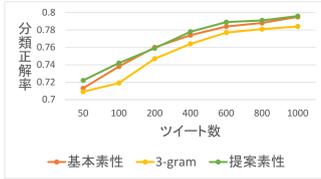


図 7 趣味-普段 (SVM)

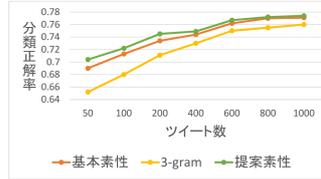


図 8 趣味-普段 (RF)

表 8 分類誤り率 (SVM)

方向性	分類誤り率
普段→公式	0.568
公式→普段	0.432
普段→裏アカ	0.504
裏アカ→普段	0.496
普段→趣味	0.488
趣味→普段	0.512

表 9 分類誤り率 (RF)

方向性	分類誤り率
普段→公式	0.653
公式→普段	0.347
普段→裏アカ	0.592
裏アカ→普段	0.408
普段→趣味	0.572
趣味→普段	0.428

表 8~9 に示す。SVM の分類誤り率では、普段用を公式アカウントと間違えている割合が高い。裏アカ-普段では、普段用を裏アカと間違える割合がわずかに高くなっている。一方で、趣味-普段では趣味用を普段用と間違えている割合が高い。RF では、3 種類すべての使い分けにおいて、普段用をそれぞれの使い分け用のアカウントに間違える割合が高くなっている。また、その割合も SVM より極端になっている。

4.3.3 素性の重要度

最後に提案素性で RF を実施した際の素性の重要度上位 5 種類を使い分け方ごとに集計した結果を表 10~12 に示す。本研究で定義づけた素性の種類は、表 1 の基本素性 5 種と表 2 の拡張素性 10 種の計 15 種類である。

3 種類の使い分け方全てにおいて名詞が最も分類の手がかりに利用された。2 番目に重要な素性は、公式-普段では記号、裏アカ-普段では単語感情極性スコアであり、趣味-普段では記号と単語感情極性スコアが同数となった。また、趣味-普段の使い分けにのみハッシュタグのキーワードが出現した。その他に、動詞や形容詞、副詞、URL のドメイン、アカウントのスクリーンネームも出現した。しかし、1 ツイートの文字数、リプライ

表 10 重要度上位 5 件中の出現素性 (公式-普段)

タイプ	出現数
名詞	65
記号	17
単語感情極性スコア	11
URL のドメイン	3
動詞	3
アカウントのスクリーンネーム	1

表 11 重要度上位 5 件中の出現素性 (裏アカ-普段)

タイプ	出現数
名詞	62
単語感情極性スコア	19
記号	9
アカウントのスクリーンネーム	4
動詞	2
形容詞	2
URL のドメイン	1
副詞	1

表 12 重要度上位 5 件中の出現素性 (趣味-普段)

タイプ	出現数
名詞	59
記号	16
単語感情極性スコア	16
アカウントのスクリーンネーム	5
ハッシュタグのキーワード	2
形容詞	1
副詞	1

の有無、リツイートの有無、ハッシュタグの有無、URL の有無、画像の有無は 60 ペアのマルチアカウントの上位 5 件中 1 つも見られなかった。

5. 考 察

5.1 分類正解率

1 アカウントにおけるツイート数が 1,000 のときの全体の分類正解率の結果について、基本素性に拡張素性を加えた提案素性の方が、文字 3-gram を素性とした比較手法より SVM, RF どちらの分類手法を用いても上回る結果になった。また、RF においては基本素性でも文字 3-gram より高い正解率で分類できている。このことから、著者特有の癖や言い回し表現より、話題や投稿形態を手がかりにした方が高い精度で分類できるといえる。また、拡張素性を追加した方が基本素性よりも高い正解率を出していることから、拡張素性は複数アカウント所持者の投稿分類に有効であるといえる。拡張素性を追加したことにより、分類が正解に変化した例を表 13 に示す。短いツイートでも、リプライ先のアカウントのスクリーンネームを頼りに分類ができていた。それ以外にも、単語感情極性スコアやハッシュタグを手がかりに分類に成功したツイートが見られた。一方で、文字 3-gram を素性とした場合でも基本素性と同等の分類正解率となった。これは提案手法で述べたように、基本素性において形態素解析器で正確に形態素を抽出できなかったこと、

表 13 拡張素性により分類が正解した事例

ツイート内容
@ユーザ名 おっ!?
@ユーザ名 なんだったー!?
@ユーザ名 ありがとう
RT @ユーザ名: タコのお刺身いただきまーす。あれ?…w https://t.co/21KxyWZb8D
意外と面白かったお (ω)
マスクしないとしにそう
#アンチ讀賣 空気嫁内坂本が、ほうむらん打ちよった…orz

1 ツイートあたりの素性の数が少ないことが原因だと考えられる。文字 3-gram では 1 ツイートあたりの抽出できる素性の数は (ツイートの文字数-2) 個であり、素性の数が多いことが分類に有効に働いたのではないかと考えられる。

各使い分け方による正解率の変化に注目すると、公式-普段が他の 2 種類の使い分けと比べて、正解率が 0.1 程度高くなっている。このことから、公式アカウントは、普段用のアカウントと使い分けが明確に異なっているといえる。基本素性が文字 3-gram の素性より趣味-普段で上回ったのは、ツイート内容や話題に依存した基本素性が趣味用のアカウントを識別するのに有効であったと言える。逆に、趣味-普段ほど差はないが、公式-普段と裏アカ-普段では文字 3-gram の方が正解率が高い。これは、文字 n-gram が文体や言い回しを反映しているからではないかと考えられる。

ツイート数による分類正解率の変化については、ツイートの数が大きくなるにつれて分類正解率が高くなっているが、上昇率は徐々に小さくなっている。これは学習データを増やしても分類器のパラメータがほとんど更新されず、一定の値に収束していることが示唆される。

提案素性で SVM 分類を行った際に、分類に失敗したツイート例を表 14 に示す。同一のハッシュタグに関するツイートが連続して誤って分類されることが見られた。これは、学習データにそのハッシュタグの使用が見られなかったため、それ以外の素性を手がかりとして予測したからだと考えられる。また、短いツイートや挨拶など、使い分けをしているアカウント双方に見られるような内容を分類するのは難しいといえる。間違えると問題がありそうな裏アカの投稿も、いくつかのツイートが誤って分類された。

5.2 分類誤り率

アカウントの方向性による分類誤り率は、SVM、RF どちらの手法も普段用のアカウントを公式アカウントに間違えてしまう割合が高くなっている。つまり、投稿内容や投稿形態において、普段用のアカウントのツイートが、公式アカウントのツイートと類似することがあるといえる。また、RF で構築した分類器は、普段用のツイートを公式、裏アカ、趣味用のアカウントに間違える傾向にある。これは、学習データで構築した決定木の分岐条件に問題があるといえる。

5.3 素性の重要度

重要度上位 5 件の内容を確認するため、計 60 ペアそれぞれ

表 14 分類に失敗した事例

ツイート内容
#なでしこ JAPAN アメリカ、守りを固めてきたなー
#なでしこ JAPAN アディショナルタイム 3 分。。
#なでしこ JAPAN カウンターこわいいいい
晴れてきたなう
ライオカイザー …
げ。フラジャイルでヒストリア入ってない…
お?
おやすみ
明けましておめでとうございます。本年も何卒よろしくお願ひいたします。
めちゃくちゃな後輩が LINE の TL で髪染めちゃった www とかほざいてるんだけど www www www www www www www
それでいつもいつも子供扱いして、社会的価値も生み出せないくせにとか知ってるわボケ

から構築された RF 分類器のうち、各使い分け方から 2 つずつ取り出してきた分類器において重要度上位 5 件の具体的な素性と種類、ジニ係数のスコアを表 15~表 20 に示す。素性中に出現する polarity とは単語感情極性スコア、username はユーザ名を表す文字列である。提案素性における RF の重要度上位 5 件は、3 種類の使い分けすべてにおいて名詞が一番多く用いられていた。名詞は人やもの、事柄を表す単語であるが、Mecab では、半角のアルファベットや記号も名詞として識別されている。素性の数で 2 番目に多かった種類は、公式-普段では記号、裏アカ-普段では感情極性スコア、趣味-普段では記号と感情極性スコアが同数であった。公式-普段で記号が重要な理由は、お知らせや告知をするときに〈お知らせ〉や【告知】など、記号を使って枠を作るためによく使われているからだを考える。裏アカ-普段で感情極性スコアが有効となった理由は、裏アカにおいて人前で言えないことや愚痴を投稿するため、感情極性がマイナスの方向に高くなったからだといえる。趣味-普段では、記号、感情極性スコアの数と同じだが、他の使い分けには出現しなかった素性の種類としてハッシュタグのキーワードが見られた。これは、趣味用のアカウントで同一の話題を共有しようとしているからではないかと考える。また、裏アカ-普段や趣味-普段では、ユーザのスクリーンネームが公式-普段よりも多く重要な素性として使われている。このことから、公式-普段と比べて特定のアカウントとやりとりが多く行われていることが示唆される。

表 15 公式-普段の重要度上位 5 件の事例 1

rank	素性	タイプ	重要度
1	】	記号	0.300154
2	@	名詞	0.225675
3	【	記号	0.193184
4	(名詞	0.186931
5	会	名詞	0.180772

表 16 公式-普段の重要度上位 5 件の事例 2

rank	素性	タイプ	重要度
1	”	記号	0.448676
2	“	記号	0.232743
3	パンケーキ	名詞	0.167904
4	.	名詞	0.160543
5	instagram.com	ドメイン	0.152067

表 17 裏アカ-普段の重要度上位 5 件の事例 1

rank	素性	タイプ	重要度
1	polarity	極性	0.256498
2	大戦	名詞	0.241360
3	RT	名詞	0.163182
4	@	名詞	0.118192
5	.	記号	0.106740

表 18 裏アカ-普段の重要度上位 5 件の事例 2

rank	素性	タイプ	重要度
1	RT	名詞	0.188067
2	フォロワー	名詞	0.162748
3	polarity	極性	0.159609
4	p	名詞	0.139800
5	moi.st	ドメイン	0.093197

表 19 趣味-普段の重要度上位 5 件の事例 1

rank	素性	タイプ	重要度
1	#	名詞	0.288300
2	polarity	極性	0.240229
3	username	ユーザ名	0.221937
4	://	名詞	0.136873
5	.	名詞	0.135000

表 20 趣味-普段の重要度上位 5 件の事例 2

rank	素性	タイプ	重要度
1	,	名詞	0.204102
2	polarity	極性	0.176315
3	in	名詞	0.118919
4	p	名詞	0.108526
5	/	名詞	0.105706

6. まとめ

本稿は、Twitter の複数アカウント所持者が投稿するアカウントを間違えないようにするために、機械学習による投稿アカウント推定手法を提案した。分類器の性能を上げるには素性の選択が重要であるが、本研究は同一著者であるが特定の目的に応じてアカウントを使い分けしていることに注目し、ツイートの投稿内容や投稿形態を表す素性を抽出した。素性の種類は、形態素解析器で取得できる基本素性とそれ以外の方法で取得できる拡張素性を組み合わせた。

提案手法の素性の有効性を検証するため、著者推定の手法で広く用いられている文字 n -gram の素性と比較した。 n の値は 3 とし、機械学習の手法である SVM と RF の 2 種類で分類実験を実施した。複数アカウント所持者のアカウントは、アカウ

ントのプロフィール欄に記載されているものとし、使い分け方を公式-普段、裏アカ-普段、趣味-普段の 3 種類に人手でラベル付けした。1 アカウントにおけるツイート件数を 1,000 としたときの 5 分割交差検定の結果、3 種類すべての使い分けにおいて、提案素性が文字 3-gram を素性としたときよりも高い平均分類正解率を出すことができた。また、全体の分類正解率の平均値は、提案手法で SVM を実施したときで、その値は 0.828 である。この結果は、形態素解析で抽出できる形態素のみを素性とした分類正解率よりも高いことから、提案素性の有効性を確認することができた。また、素性の種類としては名詞、記号、単語感情極性スコアが分類に有効であった。

今後の課題として 2 つ挙げられる。1 つ目は、分類正解率のさらなる向上である。本手法では、投稿時間を考慮した素性を導入できなかった。例えば、企業の公式アカウントと私的なアカウントを管理していた場合、勤務時間内は公式アカウントからの投稿が多くなり、勤務時間外では私的なアカウントからの投稿が多くなることが予想される。このようなことから、投稿時間を素性に組み込むことで分類正解率の向上が期待できる。2 つ目は、投稿するアカウントの使い分けを考慮した危険なツイートを検出することである。本研究は、ツイートを投稿するアカウントを間違えることによって発生するトラブルを抑制するために、ツイートから投稿するアカウントを推定する手法を提案した。しかし、実際にトラブルになるかどうかは投稿内容に依存しており、アカウントの使い分けに対応した危険なツイートを検出することは重要である。本研究では、3 種類の使い分けすべてにおいて同一の手法で投稿分類をしたが、以上のような課題に取り組む際は、使い分けの種類を限定する必要があるかもしれない。また、今回は人手で使い分け方をラベル付けしたが、マルチアカウントのツイート群から自動的にラベル付けできるようにするのも課題である。

謝 辞

本研究は JSPS 科研費 JP16H02904 の助成を受けたものです。ここに記して謝意を示します。

文 献

- [1] 総務省. 平成 27 年版情報通信白書. <http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h27/pdf/n4200000.pdf>, 2015.
- [2] モバイルマーケティングデータ研究所. 3 大ソーシャルメディアのユーザインサイト調査. https://mmdlabo.jp/investigation/detail_784.html, 2011.
- [3] 産経 WEST. 「朝日公式ツイッター」で監督を中傷、明德義塾一境の判定巡り…スマホから送信、30 分後に削除. <http://www.sankei.com/west/news/160815/wst1608150048-n1.html>, 2016.
- [4] 小林直樹. ソーシャルメディア炎上事件簿. 日経デジタルマーケティング, 2011.
- [5] 澤田みずほ, 立木茂雄. SNS における自己表現とコミュニケーション - アカウントの使い分けと放棄について -. 同志社大学社会学部卒業論文, 2013.
- [6] 木原裕二, 笹野遼平, 高村大也, 奥村学. 複数のソーシャルメディアアカウントの関連付け防止システムの構築. 情報処理学会 第 216 回自然言語処理研究会, Vol. 2014-NL216, No. 19, pp. 1-6,

- 2014.
- [7] Roy Schwartz, Oren Tsur, Ari Rappoport, Moshe Koppel. Authorship Attribution of Micro-Messages. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing(EMNLP2013), pp. 1880–1891, 2013.
 - [8] 奥野峻弥, 浅井洋樹, 山名早人. マイクロブログを対象とした100,000人レベルでの著者推定手法の提案. 第7回データ工学と情報マネジメントに関するフォーラム (DEIM2015), D8-1, 2015.
 - [9] 六瀬聡宏, 内田理, 鳥海不二夫. 大規模災害時の情報提供を目的としたツイート分類手法. 第13回情報科学技術フォーラム (FIT2014), O-010, 2014.
 - [10] Bharath Sriram, Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, Murat Demirbas. Short text classification in twitter to improve information filtering. Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 841–842, 2010.
 - [11] MeCab. <http://mecab.sourceforge.net/>
 - [12] Twitter REST API. <https://dev.twitter.com/rest/public>
 - [13] 高村大也, 乾孝司, 奥村学. スピンモデルによる単語の感情極性抽出. 情報処理学会論文誌ジャーナル, Vol. 47, No. 02, pp. 627–637, 2006.
 - [14] LIBSVM. <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>
 - [15] Scikit-learn. <http://scikit-learn.org/stable/>
 - [16] 松浦司, 金田康正, “n-gram 分布を用いた近代日本語小説文の著者推定,” 情報処理学会研究報告自然言語処理, Vol. 99, No. 95, pp. 31–38, 1999.