令和 3 年度

博士論文

深層ニューラルネットワークを用いた Twitter ユーザの興味推定に関する研究

板東浩二

徳島大学大学院先端技術科学教育部

システム創生工学専攻　博士後期課程3年

内容

図目次

表目次

第 1 章　はじめに

1.1　研究背景

近年，ソーシャルメディアを通じたユーザ同士のコミュニケーションが大きく増加している．同じ趣味の仲間を見つけるために，Twitter や Facebook 等のソーシャルネットワークサービス（social networking services; SNS）を利用することは一般的な手法となりつつある．しかし，誰もが SNS を通じて気軽に情報発信・コミュニケーションをできるようになった結果，情報量は膨大なものとなり，1 件 1 件のユーザを観察しながら同じ趣味の仲間を探索していくことは困難となっている．

簡便な方法として，Twpro[1]のように，趣味，性別，年齢，場所，仕事など，似たような属性を持つユーザをプロフィール検索できる Web サービスも存在する．これらは共通の趣味を持つユーザを発見する上で非常に有用なものとなっているが，検索の結果として得られるユーザのプロフィール情報と，個々のツイートは常に対応しているとは限らない．たとえば，「映画鑑賞が趣味」とプロフィール情報に記載されてはいたとしても，多くのツイートは日常生活に関するものであり，映画鑑賞のツイートはわずかな部分しかないなど検索目的に沿わないケースがあげられる．また，仕事や趣味，年齢が変わってもプロフィール情報が更新されず，ユーザ属性として参考にならないケースも存在する．さらに，匿名性を確保するためにプロフィール情報を意図的に少なくするユーザや，虚偽の内容が記載されているケースもある．一方でプロフィールとツイート内容が一貫して一致するケースは広告・広報用のアカウントの場合が多く，必ずしも適切なユーザとは限らない．

プロフィールによるユーザ探索は有効な場合も多いが，問題となる要因も多数あり，同じ趣味を持つ仲間を探索する上で最も良い結果を得られるとは限らない手法となっている．プロフィールを用いない探索方法として，話題になっているトレンドや，キーワ

[1] https://twpro.jp/

ードを利用したリアルタイムのツイート検索[2]など現在時点での情報を収集するサービスも存在している．最新の情報を得られるため，ユーザ探索の精度は高いが，ある一瞬のツイートを切り取ってもユーザの総合的な性質とは合致ない可能性も高い．これらのサービスはあくまでビッグデータ解析にとどまっており，ユーザ個人での性質や興味などを考慮したものとはなっていない．

1.2 研究目的

本研究では，プロフィール情報を使わず，Twitter の連続したツイートからユーザの興味（趣味）を推定する手法を提案する．1 つのツイート自体には多くの情報が含まれず，一部のツイートはユーザの趣味に関連しないケースが多いため，信頼性を高めるために複数の連続したツイートから特徴を抽出する．これにより，一過性の話題や一般的に，趣味に関連するキーワードのほとんどは名詞であるため，固有名詞に対応できる必要がある．固有名詞は一般辞書に含まれていないため，提案手法は意味/文脈を適切に処理できる単語分散表現を抽出することにより，カテゴリ分類の多様性を拡張する．

近年，ディープラーニングを用いた研究が多く，大量のテキストデータを取得できる現状において，ディープラーニングはテキスト分類においても効果的である．本手法では，連続する複数ツイートに対し，ツイート単位で分散表現の平均ベクトルを抽出し，そのベクトルについて時系列情報を失わずに学習できる，深層学習の一手法である再帰ニューラルネットワークと畳み込みニューラルネットワークを用いる．

[2] https://search.yahoo.co.jp/realtime

1.3 本論文の構成

本論文では第 2 章で関連研究について述べる. 第 3 章では, 提案手法である RNN と CNN を用いた分析, その要素技術について説明する. 第 4 章では, 提案手法を用いた評価実験と実験結果の考察について述べる. 第 5 章では, まとめと今後の課題について述べる.

第 2 章　関連研究

ソーシャルメディアからのユーザ興味推定は多くの言語圏で，さまざまな手法の研究が実施されている．その 1 つに，ソーシャルメディア上のビッグデータを対象とした分析に関する研究があげられる[1-9]．Sloan ら[1]はイギリスの Twitter プロフィール情報から人口統計データを自動的に抽出する実験を行い，Ren ら[2]は SNS のビッグデータから感情分析を行うための方法を提案した．Magumba ら[3]は深層ニューラルネットワークを活用し Twitter のツイート情報から病名を抽出した．これらは，大規模なデータから効率よく人間が理解しやすい形式で情報を抽出するかに着目している．計算機資源の充実や，計算アルゴリズムの高度化により，ある程度のデータ量までなら分析も可能となった．

しかし，近年ではビッグデータからの分析だけでなく，個人の行動に注目した分析が重要となっている．たとえば，映像配信サービスにおいて視聴者の動向を把握し，人気のコンテンツをユーザに薦めることは一般的な手法として既に確立されている．現在重要となっているのは，視聴者個人の視聴傾向から興味を分析し，ユーザひとりひとりが興味を持つコンテンツを薦めることである．そこで，本章ではビッグデータではなく個人の行動を分析した研究に注目する．ソーシャルメディアの情報からユーザの興味を分析し，分類した研究について述べる．

2.1　ユーザが投稿した文章の要約による分類

Forss ら[10]は人々のソーシャルメディアのプロフィールを分析し，趣味や関心事の情報を抽出する方法を検討した．プロフィール分析によるユーザ興味抽出のベースライン手法として，ヒューリスティックなルールと TF-IDF を用いたものを提案し，英語圏ユーザの分析を実施した．ユーザの興味のある単語抽出のためのベースラインシステムに，関連するコンテンツの範囲の制限，名前付きエンティティの抽出，スコアの低い趣味や関心事を特定するための事前定義された辞書の使用，多言語のコンテンツを処理するための機械翻訳の使用などの機能を追加している．

結果から，Fross らはソーシャルメディアの分析では，「抽出に失敗したり，認識されない単語」が他の単語と比べて重要と誤認される問題があると述べた．そのためにも複数言語に対応しておくことは重要性が高いとの知見を示している．

また，文に基づいた要約よりも，キーワード抽出の方がソーシャルメディアコンテンツの分析に利用しやすいという点を述べている．キーワード抽出を事前定義された辞書や名前付き実体認識と組み合わせることで，より広い範囲をカバーすることができるためであり，キーワード抽出の優位性を説明した．

2.2 マイクロブログにおける興味タグを用いたユーザの分類

ユーザが公開したメッセージには，その個人の興味が反映されることが一般的である．Yuら[11]は中国語のマイクロブログの特徴に注目して，テキストの特徴量をもとにクラスタリングと分類アルゴリズムを組み合わせ，ユーザの興味に合致した興味タグを抽出した．この手法では，クラスタリングと分類を組み合わせた戦略を採用して，ユーザをモデル化している．以下2つのアルゴリズムを組み合わせることで，ユーザの関心に沿った興味タグを抽出することができたと述べている．

①Weibo のみに出現するストップワードの除去と，そのテキストのベクトル化空間表現に焦点を当てた WUKE アルゴリズム

②Weibo のテキスト・ベクトルを扱う際に，クラスタリングと分類を組み合わせることで，ユーザの関心事であるタグを取得する WUTE アルゴリズム

マイクロブログのテキストは，他のテキストとは異なり，特殊な句読点や顔文字の他に，マイクロブログにしか登場しない単語が含まれている．そのため，本手法はマイクロブログに特化していることではあるが，クラスタリングと分類を組み合わせた方法で興味タグを抽出することで，ユーザの興味タグの分類精度が向上する可能性があると述べている．

2.3 感情とプロフィールに着目した関心事によるユーザの分類

Lewenberg ら[12]はツイートからのユーザ興味分析のためには感情分析が重要だとし，感情を元に興味を推定する方法を提案している．ユーザのプロフィール及びフォローしているアカウントやフォローすべきアカウントなどの関心事と関連付け，認識されたユーザの属性と興味の間の相関関係について調べた．

この手法では，ツイートの文章から取得されたユーザの感情を Ekman の 6 つの基本的な感情（Anger, Disgust, Hear, Joy, Sadness, Surprise : Ekman's six basic emotions）に当てはめて分類している．この手法を実現するため，Lewenberg らはユーザが感情を表現する傾向と，様々な分野への興味の度合いとの間に統計的に有意な関係があるかどうかを検証したのち，ユーザが表現した感情に基づいて，ユーザの興味のある分野に関する予測を行った．具体的には，ユーザが表現した各感情のスコアをロジスティック回帰モデルに取り込み，Ekman の 6 つの感情のそれぞれに対するユーザの感情スコア（割合スコアまたは分布スコアのいずれか）を取るロジスティック回帰モデルを構築し，最終的にこの予測モデルの品質を評価している．

このモデルは，ユーザの興味とその表現された感情に関する予測をある程度可能にしたものの，いくつかの問題を孕んでいる．抽出された関心度が第3者によるプロフィールの評価をもとに行われているため，本来の関心と合致しているかは不明である．同様に，感情スコアを決定する方法は機械学習モデルに基づいており，完全に正確ではない可能性がある．特に，特定のトピックに興味を持つ人が特定の感情パターンを採用する傾向があるのか，あるいは様々な感情を経験する人が特定の関心分野に惹かれる傾向があるのかは不明である．

Lewenberg らは予測モデルの性能は良いものの，異なるトピックに興味を持つユーザが表現する感情のパターンが似ているケースがあると述べている．より詳細な感情のリストを用いることで，ユーザの関心事の予測性能を向上させることができるかもしれないと展望を述べた．このように，元からあるモデルにユーザ属性を当てはめる場合，より詳細に分類されたモデルでなければ，ユーザ特徴のディティールは失われてしまう．

2.4 フォロー関係や土地情報に注視したユーザの分類

フォロー関係から，ユーザの特定の情報を抽出しようとする試みは多く行われている[13-16]．たとえば Agarwal ら[13]は友人関係のネットワークから趣味を推定する方法を実験し，特定の趣味に対しては有効な分析ができると示した．筆者らは自己申告した友人関係と，SNS で定義された有限のリストから自分で選んだ特徴や趣味を含む新しいデータセット「TravelSite」を導入し，グラフ変換アルゴリズムを提案している．このアルゴリズムでは友人ネットワークの関連性，ひいては McPherson ら[17]のホモフィリー仮説妥当性について検討されている．結果，自己申告された友人ネットワークから，オンライン・ソーシャル・ネットワークに参加している人々の趣味を予測できることを示した．

また，Mangal ら[18]は，これまでに述べた手法に加え更にユーザの位置情報やツイートを考慮する手法が用いられている． 特定のトピック（エンターテイメント，政治，スポーツ，テクノロジー，ビジネスなど）に関するツイートに基づき，特定の場所にいるユーザの関心との関連性を調べている．ツイートの感情を分析し，その後，トピックに応じてツイートを分類することで，ユーザを分類し，興味を抽出した．

2.5 属性と辞書によるユーザの分類

ユーザの興味推定において，ユーザの属性を求めることは重要な要素となる[19,20]．Ritter ら[19]は Twitter におけるユーザ属性推定のためのフレームワークを提案し，特に 3 つの有用なプロフィールの属性として教育，仕事，配偶者という属性に分類してそれぞれ正しく分類できたかどうかの実験を行なった．Ritter らは経歴や趣味などの属性を共有している人ほどソーシャルメディアで友達になる確率が高いというホモフィリー仮説を用い，特徴ベクトルに含まれる属性の関連性に着目した．教育，仕事，配偶者それぞれの属性が異なった関連を持つため，この 3 つの属性について，3 つの分類器を別々に学習している．結果，弱教師つき学習手法を実現した．

この際，文章から抽出したテキストのみを興味として抽出した場合と，ブラウザ中の言語を抽出して学習し，カテゴリと関連語に分類し，提供している辞書サービス

（never-ending language learning; NELL） を用いた抽出を行なった場合について比較している. 結果, NELL を用いた学習は教育の属性については非常に有効な結果を残している. これは教育ネットワークが仕事ネットワークよりも強いホモソフィー特性を示していることに影響を受けている. このように NELL は教育属性の推定では大きな利益を得ることができるが, 仕事推論では限定的であった. これは教育に関する言及は, 通常, 宿題, 試験, 勉強, カフェテリア, 本などの明確な証拠と関連しているが, 仕事については, 語彙が通常, 異なるタイプの仕事に特有であるため, 状況ははるかに複雑であることに起因している. この手法は膨大な教師データを必要としないが, 分類に用いる辞書の影響を大きく受けてしまうことになる. 推論性能を高めるためには, より豊かな特徴空間を取り入れることが必要となっている.

このように, プロフィール情報を用いてユーザの興味を推定する方法は多く検討されている. しかし, Twitter に関して言えば, この手法は最適とは言い難い. Twitter に登録されているプロフィール情報のほとんどは, 趣味, 年齢, または仕事が変わってもユーザによって情報が更新されないため, そもそも分類に用いる情報が誤っている可能性があるためである.

2.6 ツイート情報に注目したユーザの分類

一方で, ユーザの興味をツイートから収集したトピックごとに分析する手法も検討されている. 渡邊ら[21]は潜在的ディリクレ配分法を用いてトピックを分析し, 更に協調フィルタリングを使って興味推定を実施した. 上條ら[22]はアンケート調査とユーザデータの解析を使ってユーザの性格特性を推定するモデルを作成している. Twitter でのユーザの日常のツイートを分析する[23,24]ことにより, ユーザの性格を推定する研究もある. また, 年齢, 性別, 職業などのユーザのプロフィールや属性に焦点を当てた多くの研究がある[22-26]. 加藤ら[24]は, Twitter でのユーザの属性と習慣的な行動を推定した. 彼らの方法では, 投稿されたコンテンツとユーザのプロフィールテキストだけでなく, ユーザのライフスタイル情報も使用した. 市販製品およびテレビ番組に関する意

見を抽出するために，池田ら[25]は，Twitter に投稿された意見を分析して，年齢，性別，地域などのユーザのプロフィールを推定した．

2.7 サポートベクターマシンによるユーザの属性分類

Rao ら[26]は，Twitter ユーザに対し，性別，年齢，地域，政治的指向の 4 種類のユーザの属性分類方法に基づいて，サポートベクターマシン（support vector machine; SVM）をベースとした分類アルゴリズムを用いて実験を行った．SVM は分類タスクにおいて高い性能を発揮する機会学習のモデルであり，Rao らは SVM を発展させた機械学習でユーザの属性分類を実施した．Twitter では文法や単語が正確ではない，くだけた文章が多いが，提案手法を用いることで他のベースライン方法よりも優れたパフォーマンスを達成している．

この研究では機械学習が有効であること，ユーザ属性となるカテゴリを絞って分類することで高い精度が出せることを示している．ただし，その方法はユーザの属性そのものを推定するものであり，ユーザの興味や趣味を推定するものではない．

2.8 連続したツイート情報を元にユーザ興味を推定

本章で述べたように，プロフィール情報やツイート情報からユーザの属性や興味（趣味）を抽出する試みは多く行われている．しかしこれらの研究の多くは，取得したユーザのツイートを 1 つのデータ集合として扱い，ツイートの順序（時系列情報）を考慮していない．そこで，本稿では時系列データを用いたユーザの連続ツイート情報を扱い，強力な機械学習の手法である深層ニューラルネットワークを使ったユーザの興味推定を提案手法として 3 章で述べる

第 3 章　深層ニューラルネットワークを用いた興味推定

3.1　提案手法の概要

本研究では，即時性の高い Twitter を利用することで，変動のある属性情報である個人の興味（趣味）を抽出することに焦点をあてる．大規模データをいきなり扱うのではなく，ある一定の条件のもと収集したデータセットのなかでの実験を行うことで提案手法の有効性を検証する．

提案手法の概要は，図 1 のようになる．提案手法では，ユーザの発言を時系列順に取得し，S ツイートずつずらしながら N ツイート系列を作成する．その際，ツイートが重複することを許す．得られたツイート系列に対し，ツイートを形態素解析により単語単位に分解し，単語列に変換する．単語列中の単語に対し，学習済み単語分散表現のモデルを参照し D 次元の単語分散表現ベクトルを抽出する．単語分散表現ベクトルの平均ベクトルをツイート単位で作成することで，1 ツイートにつき $N \times D$ の行列を得る．

これを素性 X として，推定対象 Y(趣味カテゴリ)とともに再帰型ニューラルネットワークを用いて学習させることで，N ツイート系列からの趣味カテゴリベクトルの推定をおこなう趣味カテゴリ推定器を構築する．1 ユーザにつき，複数の推定結果（趣味カテゴリベクトル）が得られることから，それらの平均値を最終的な推定結果として，評価をおこなう．

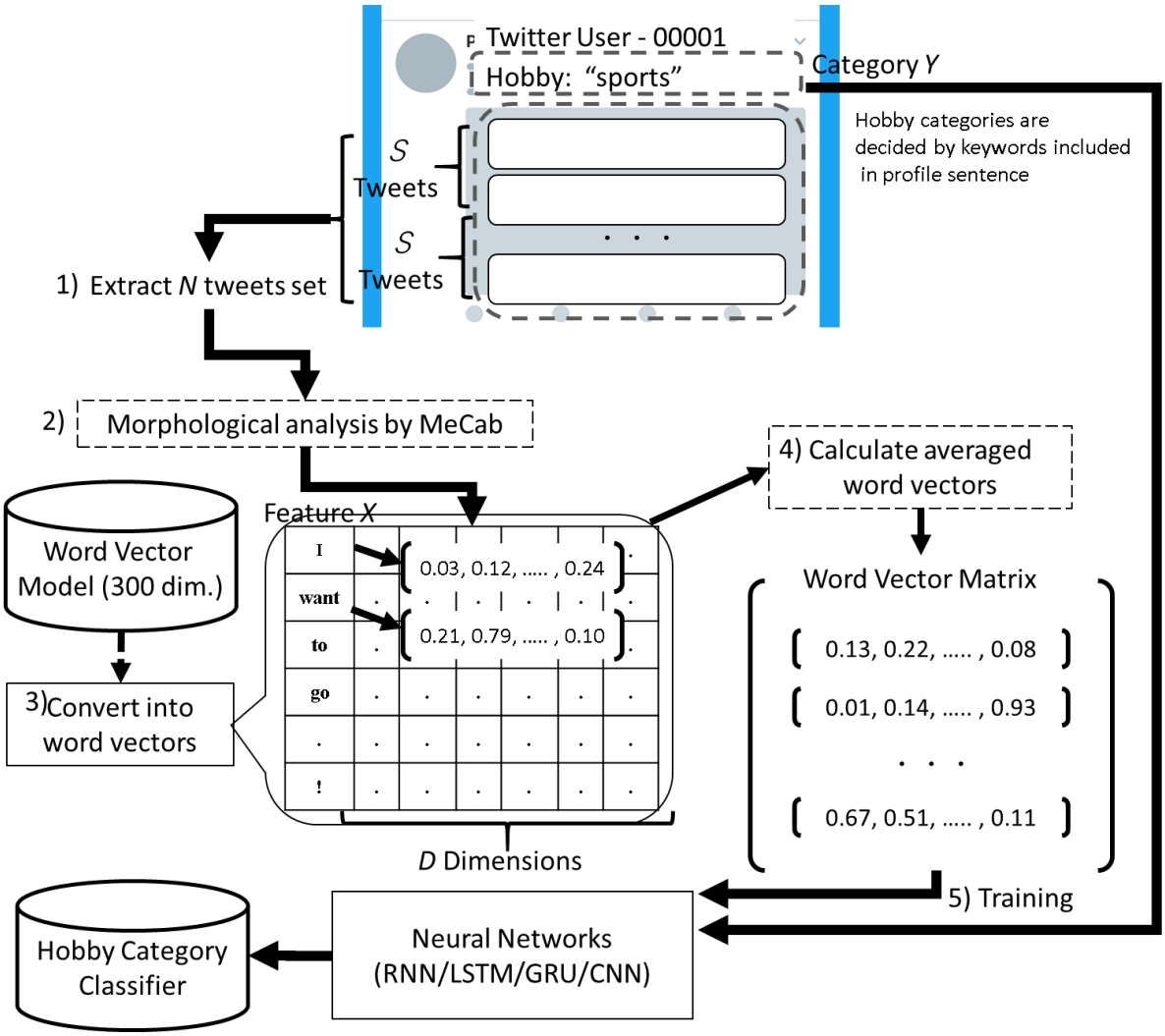


図 1 時系列ツイートの解析と学習方法

3.2 ニューラルネットワーク

ニューラルネットワークは人間の神経細胞を基にし，その細胞同士をネットワーク構造でつないでデータ処理を行うモデルである．基本として入力層・中間層（隠れ層とも呼ばれる）・出力層を持ち，データを入力層に流し，中間層で計算を行い，出力層に結果を出す．この層が多いものを深層ニューラルネットワークと呼ぶ．

ニューラルネットワークによる学習としては「教師なし」「教師あり」があり，「教師なし」の代表例としては自己組織化マップ（self organizing map; SOM）がある．SOM は人間の脳の仕組みをモデル化したもので，ニューラルネットワークにさまざまなデータを入力することにより，データ間の関係性を低次元で表現することができる．出力層では出力する次元を指定可能となっており，高次元データを 2 次元にマッピングすることで視覚的なデータの分析に役立つ．「教師なし」学習の多くはデータの分析や傾向の調査に用いられるものであり，より高度な学習成果を期待する場合は「教師あり」が多く採用される．

「教師あり」の学習は出力層が教師データに近づくように中間層の重みを調整していく方法が一般的である．最も基本となるモデルは入力層→中間層→出力層と順に情報が流れていき，逆には流れないことから順伝播型ネットワークと呼ばれる．図 3 に基本的な順伝播型ニューラルネットワークを示す．

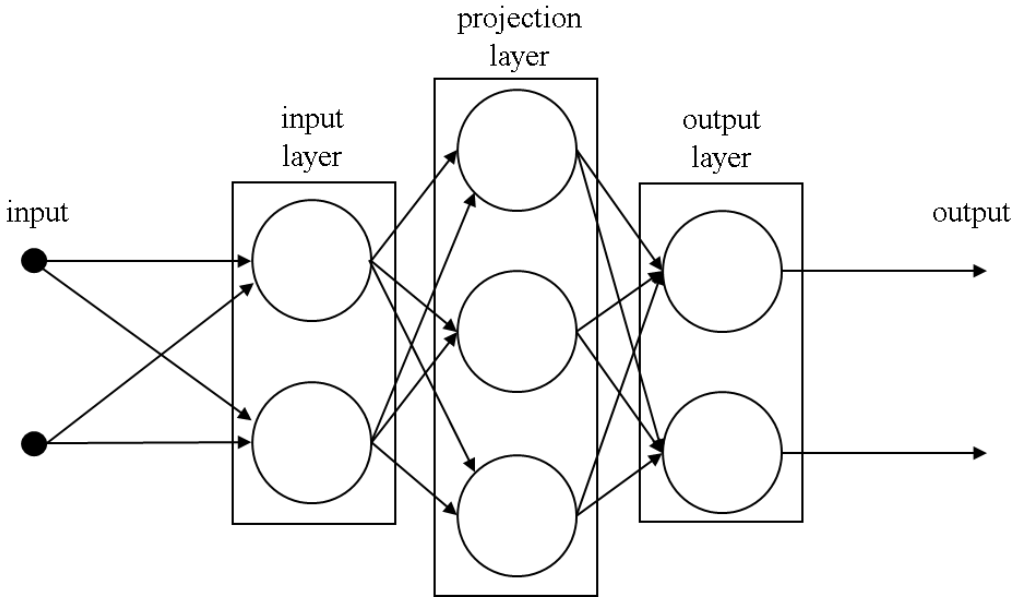


図 2 基本的な順伝播型ニューラルネットワーク

再帰型ニューラルネットワーク（recurrent neural networks; RNN）はニューラルネットワーク内に再帰的な構造を持つ手法である．主に時系列ごとに変化するデータの学習に用いられる．最も基本的な順伝播型ネットワークと比較すると，データを入力・処理・出力をしていく処理フローは変わらないが，途中にループ構造がありネットワーク内を循環する動きをする．時間ごとにデータを入力していき，ある時間における中間層からの出力が別の時間の中間層に引き渡されるという特徴がある．これにより，自然言語や音声など時系列データに対して高い学習効果を期待することができる．本研究では，ツイート中の単語の並びよりも，ツイートの並びに意味があると考える．Twitterは即時性の高いメディアであるため，1 つの出来事を複数に分けて投稿することが多い．そのため，連続するツイートは関連性が高いと考えられる．

一方で，単語の並びにも意味はあるが，Twitter で投稿される内容の多くがきわめて口語的であり，単語やフレーズのみの投稿も多いため，単語順よりも，同じツイート内にどのような単語が共起しているかのほうが重要と考えられる．RNN は時系列データを扱う上で記憶のように学習過程を扱う．しかし，ネットワークが複雑になると勾配消失や勾配爆発等，学習が上手くいかないケースも出現する．問題を解決する手法として，

古い記憶（大きく昔の時間データ）を打ち切って学習する方法があるが，過去の記憶を十分に活用できないというデメリットを持つ．これらの問題の解法として，長期記憶・短期記憶のように分けて行う学習（long short-term memory; LSTM）[29]がある．この手法は RNN の発展系となるもので，RNN という種別の中では主流なものとして用いられている．

しかし，LSTM は強力な手法であるが，ネットワークが複雑になるという問題点もある．ゲート付き回帰型ユニット（gated recurrent unit; GRU）と呼ばれる手法も RNN の発展系の一つであり，LSTM より比較的簡単な構成となっている．LSTM より計算量は小さく一部の分野では同等の性能を出すことができるが，総合した性能は LSTM の方が上になることが多い．本研究では RNN の発展形として LSTM, GRU の 2 つの手法を利用する．

さらに，時系列は学習しないが，隣接するツイート間の関係を学習できる手法として，畳み込みニューラルネットワーク（convolutional neural network; CNN）も用いる．実験に使用した RNN, LSTM, GRU, CNN のネットワーク構造について図 5〜8 に示す．出力層の活性化関数に Softmax，最適化アルゴリズムに Adam を用いた．また，過学習を防ぐため，各レイヤーの出力の際に Dropout 率を設定している．

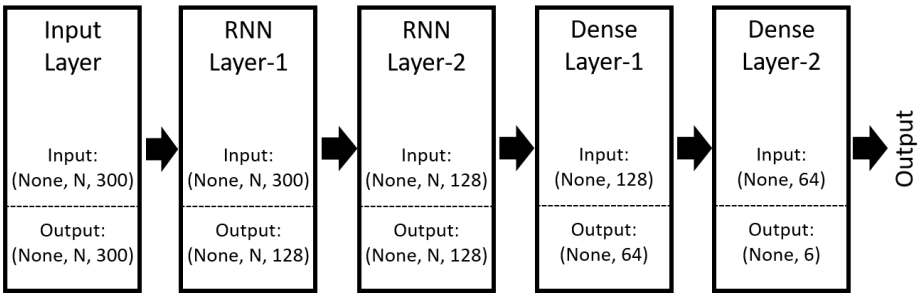


図 3 RNN ネットワーク構造

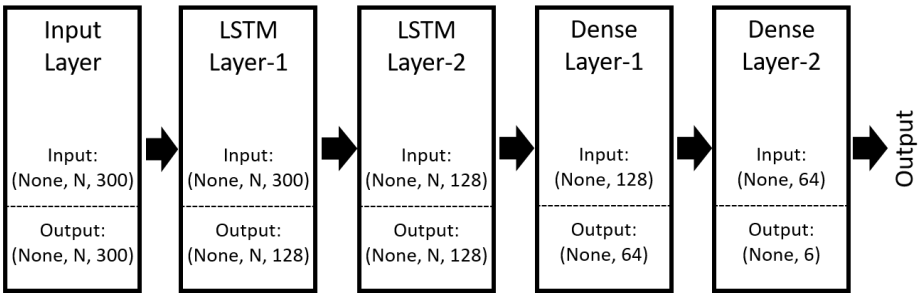


図 4 LSTM ネットワーク構造

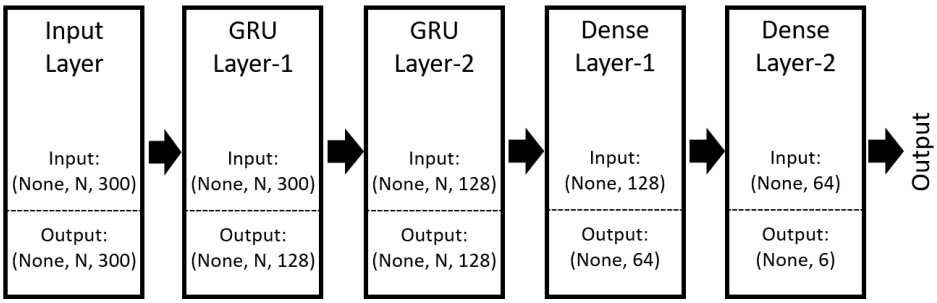


図 5 GRU ネットワーク構造

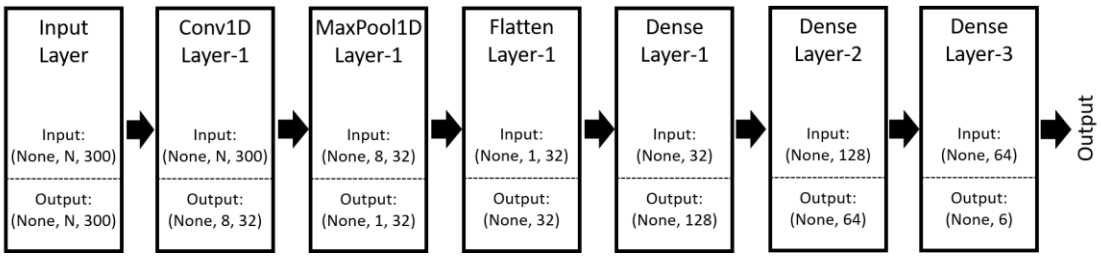


図 6 CNN ネットワーク構造

各ネットワークの構築には Keras 2.1.1[3]を用い，Tensorflow1.4.0[4]をバックエンドフレームワークとして用いた．Epoch 数の最大を 10 とし，損失関数に MSE(Mean Square Error)を用いた．また，Loss 値が改善されない場合には学習を打ち切る Early Stopping 法を用いた．

各ニューラルネットワークには，ツイートから計算した単語の平均ベクトルの列を入力する．入力されるベクトル列（行列）を式(2)に示す．$v_1, v_2, \cdots, v_N$は，それぞれ単語平均ベクトルを示している．

$$Input_i = (v_1, v_2, \cdots, v_N\ ) \qquad (2)$$

3.3 単語分散表現

自然言語を解析するためには単語をデータとして認識して定量化することが必須となる．単語分散表現は 2013 年頃 Google の研究者が Word2vec[5]という手法を公開したことにより，爆発的に広がった有力な自然言語処理の手法である．単語分散表現では，単語を多次元ベクトルとして表現することで，単語間の関係性をとらえることが可能となっている．代表的な例としては，king – man + woman = queen という四則演算の計算による単語間の表現がある．

[3] https://keras.io/

[4] https://www.tensorflow.org/

[5] https://github.com/dav/word2vec

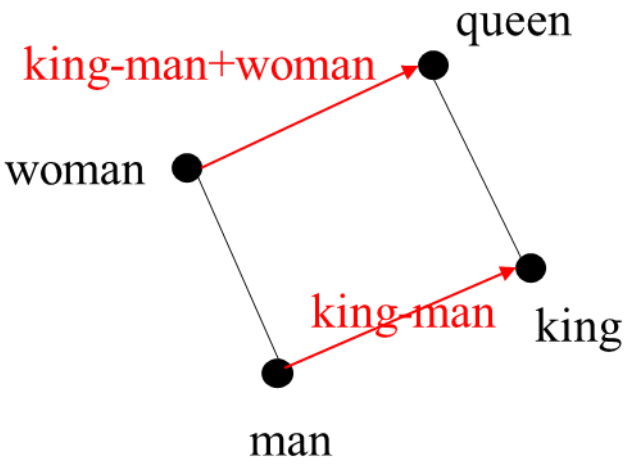


図 7 単語分散表現の代表例

図 2 で示すように man, king, woman, queen という単語をベクトルとして表現することでそれぞれの関係性を計算可能とする. king-man のベクトルに woman のベクトルを足すことで queen というベクトルを表現することができる. このように単語分散表現を用いることで単語を定量的に表現しつつ, その関係性を計算する.

Word2vec は, テキストコーパスから単語分散表現を学習するアルゴリズム/ツールであり, 意味の似ている単語同士が類似するベクトルを持つような学習を可能とするため, テキスト分類, 意味解析や機械翻訳などに広く応用されている. 単語分散表現のモデルには, Word2vec で用いられる Skip-gram Model や, CBOW(Continuous Bag of Words)などがある. Skip-gram Model はある単語を与えたときにその周辺語となるものを推定するためのニューラルネットワークの学習モデルで, CBOW は周辺語から対象となる単語を推定するニューラルネットワークの学習モデルである. 連続した 5 つの単語w_1, w_2, w_3, w_4, w_5を持つコンテクストを例として、Skip-gram の概要を図 8, CBOW の概要を図 9 に示す.

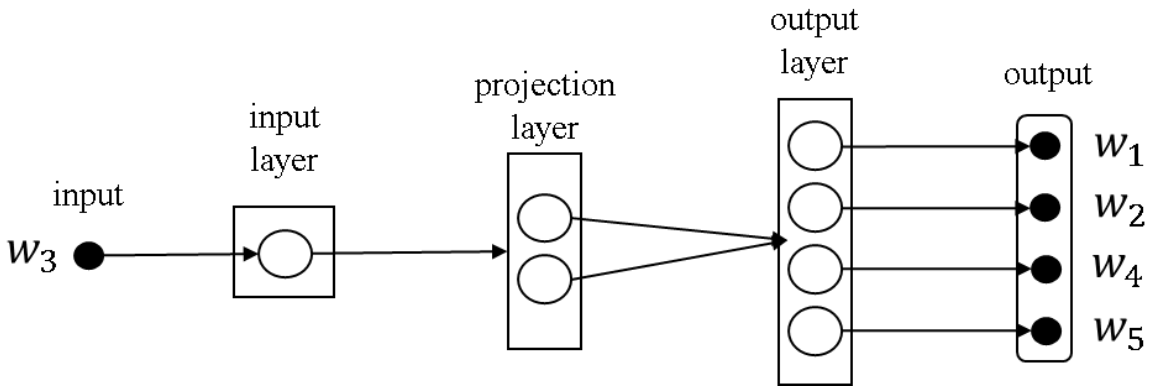


図 8 Skip-gram の概要

Skip-gram では周辺語を求めるためにw_3を入力として与えたとき，その結果の周辺語としてw_1，w_2，w_4，w_5を得ることができるモデルである．

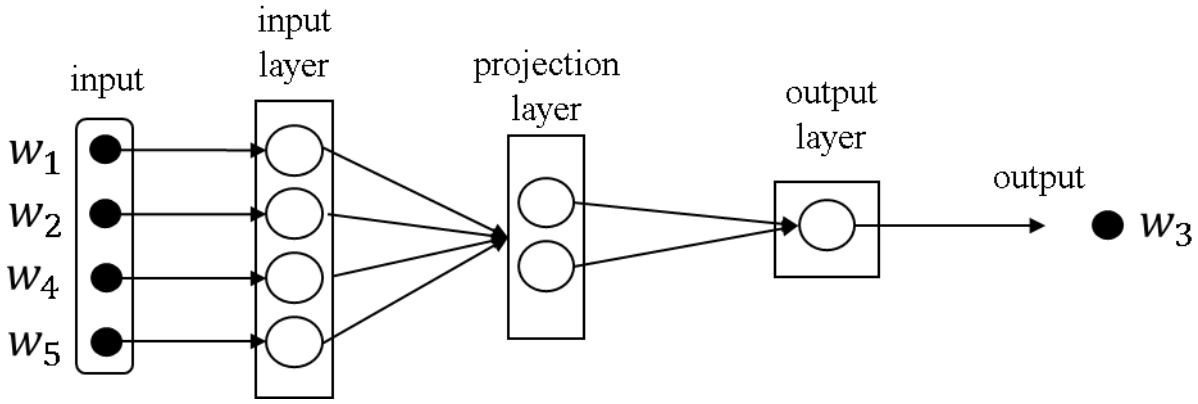


図 9 CBOW モデルの概要

CBOW では周辺語から間にある単語を求めるため，w_1，w_2，w_4，w_5を入力として与えたとき，その結果としてw_3という間の単語を得ることができるモデルである．一般的には Skip-gram の方が CBOW よりも精度の良い結果が得られるとされているが，計算量が多く時間がかかるという問題がある．

Word2vec の他にも，GloVe[28]のような手法もあり，Word2vec よりも学習が速く，精度が高く，かつ，小さなコーパスでも動作可能といったメリットがある．これは，共起行列を学習に加えることにより，精度の良い初期値が得られるからである．しかしながら，どの方法が良いかはタスクや対象となるデータに依存する．

本研究では，fastText[6]により事前学習されたモデル[29]を用いる．fastText は Facebook 社にて開発された自然言語処理のためのライブラリである．fastText でも Skip-gram と CBOW のモデルが使われているが，改良が加えられており Word2vec よりも高速で精度の高い結果を得ることができる．

また，fastText の特徴としてサブワードを学習に使う点がある．たとえば「neural」「network」という 2 つの単語に近しい関係性があるとした場合，fastText では「neural」と「network」が類似している，というように学習を進めていく．一方で fastText では n-gram で単語を解釈して関係性を学習する．tri-gram の場合であれば「ne」「neu」「eur」「ura」「ral」「al」というように分割したサブワードと呼ぶ単位で近しい単語を学習していく．これにより，類似する文字列のベクトルが近しくなるように学習することができるため，単語の表記ゆれや未知表現に頑健という特徴を持つ．

提案手法では，ツイートを日本語形態素解析器 MeCab[7]により形態素解析して得られた単語列から，単語分散表現の集合を抽出し，その平均ベクトルを生成することで 1 ツイートごとの素性を得る．このベクトルモデルの次元は 300 次元であり，日本語 Wikipedia 記事を学習データとして用いている．

式(1)は，平均ベクトル計算式である．v_xがツイートxの平均ベクトル，$|W_x|$はツイートxに含まれる単語の数，wv_x^iは，ツイートxに含まれる単語W_iの単語ベクトルを示す．

$$v_x = \frac{1}{|W_x|}\sum_{i=1}^{|W_x|} wv_x^i \qquad (1)$$

[6] https://github.com/facebookresearch/fastText/blob/master/pretrainedvectors.md

[7] http://taku910.github.io/mecab/

第 4 章　評価実験

4.1　実験に使用するツイートの収集

ユーザアカウントの趣味情報の収集方法と，それに対応したツイートの収集方法について述べる．まず，Twpro の Web サイトより「趣味」で検索する際の大カテゴリ 12 についてアカウント情報を取得する．アカウント情報の取得には，Twpro API[8]を用いた．Twpro を用いて，ある任意のキーワードに合致するユーザを検索する例を図 10 に示す．

Ex.) query=”クッキング”, number of display=10

https://twpro.jp/1/search?q=cooking&num=10

Result:


```
{
  "total": 11876,
  "count": 10,
  "users": [
    {
      "created_at": 1378992193,
      "description": "人気レシピや簡単レシピ美容レシピなど厳選して紹介していきます。
      め知識も増えるよ♪良かったらフォロー&RTお願いします☆",
      "followers_count": 48712,
      "friends_count": 4,
      "id": 1857411446,
      "lang": "ja",
      "listed_count": 466,
      "favourites_count": 0,
      "location": "",
      "name": "簡単クッキング",
      "profile_image_url": "http://pbs.twimg.com/profile_images/3788000004684611511
      c7490eb_normal.jpeg",
      "profile_banner_url": "https://pbs.twimg.com/profile_banners/1857411446/1378
      "protected": null,
      "screen_name": "Cooking_f",
      "statuses_count": 10975,
      "time_zone": "Irkutsk",
      "url": null,
      "verified": null
    },
```


図　10　アカウント情報の収集

[8] https://twpro.jp/doc/api/search

続いて，各アカウントに対し，Twitter API[9]を用いてタイムライン情報を取得する．Twitter API を Python から使用するライブラリ Tweepy[10]を用いて収集した．1 アカウントごとに約 20 ツイートを取得した．

4.2 実験方法

収集したアカウントのうち，ユーザ数が 200 件以上得られた趣味カテゴリを分類対象とする．また，一つのカテゴリ内でさらに詳細に分類可能なものについては，特徴が分散してしまうと考えらえるため，対象から除外した．music, gourmet, craft, game, art, sports の各カテゴリからランダムで 200 アカウントを選択し，分類対象のユーザとした．1 つの入力に用いる連続するツイート数は，3〜15 を試した．ツイート数を 3 以上とした理由として，CNN は前後のベクトルの畳み込みを行うためのフィルタを用いるが，フィルタのサイズを 3 としているためである．

また，ベースライン手法としてユーザの取得ツイートすべてから Bag of Words ベクトルを生成して特徴ベクトルとする手法を用いる．ベクトルの各次元は単語，値は出現頻度とした．機械学習手法には，Random Forests と SVM を用いる．単語の種類が多く，特徴次元数が膨大になることを防ぐため，χ^2値による特徴選択を用いた．

実験結果の評価には，5 分割交差検証(5-fold cross validation)を用いて，評価指標に正確度（Accuracy）,精度（Precision）,再現率（Recall），F1 スコア を用いる．数式(3)-(6) に，それぞれの計算式を示す．

$$\text{Accuracy}(\%) = \frac{1}{5} \times \sum_{i=1}^{5} \frac{C_i}{T_i} \times 100 \tag{3}$$

$$\text{Precision}_x(\%) = \frac{1}{5} \times \sum_{i=1}^{5} \frac{C_i^x}{pred_i^x} \times 100 \tag{4}$$

[9] https://apps.twitter.com/

[10] http://www.tweepy.org/

$$\text{Recall}_x(\%) = \frac{1}{5} \times \sum_{i=1}^{5} \frac{C_i^x}{true_i^x} \times 100 \quad (5)$$

$$\text{F1-score}_x = \frac{Precision_x \times Recall_x \times 2}{Precision_x + Recall_x} \quad (6)$$

式(3), 式(4), 式(5)における i は, 分割したデータセットの番号を示す. 式(3)において C_iはデータセット i における趣味カテゴリの推定に正解したアカウント数, T_iはデータセット i におけるユーザアカウント数を示す. 式(4)における$pred_i^x$は, データセット i において趣味カテゴリ x と推定されたユーザアカウント数, C_i^xはデータセット i における趣味カテゴリ x の推定に正解したアカウント数, 式(5)における$true_i^x$は, データセット i において趣味カテゴリ x であるユーザアカウント数を示す.

使用したデータの概要を表 1 に示す. ツイートに対して, アカウント ID や記号列, リンク URL などの情報は分類に不要であるため, 正規表現によりあらかじめ除去した後に形態素解析をおこなった. 具体的な正規表現については図 11 に示す. また, 学習データと評価データの概要を図 12 に示す.

表 1 データセットの詳細

Label	**# of words**	**# of uniq. words**	**# of ツイート s**
sports	98054	12738	3389
art	91981	12583	3136
music	90752	13270	3167
game	87686	12368	3097
groumet	86621	12545	2928
craft	65750	10917	2372

- .*(#|＃)[^¥s]+){2,}
- amzn¥.|amazon¥.|rakuten¥.|a¥.r10|appstore¥.|itn¥.|dmm¥.|info¥-zero¥.)
- .*(@|＠)[^¥s]+){2,}
- .*(拡散|募集|相互|支援|sougo|rt|RT|RT|follow|フォロ|ﾌｫﾛ|100|100|％|%)){2,}
- (RT)*@[a-zA-Z0-9¥_]+
- https?://[¥w/:%#¥$&¥?¥(¥)~¥.=¥+¥-…]+
- RT¥s?[¥w/:@¥_%#¥$&¥?¥(¥)~¥.=¥+¥-…]+
- [#]([¥w一-龠ぁ-んァ-ヴーa-z]+)

図 11 不要情報除去の正規表現

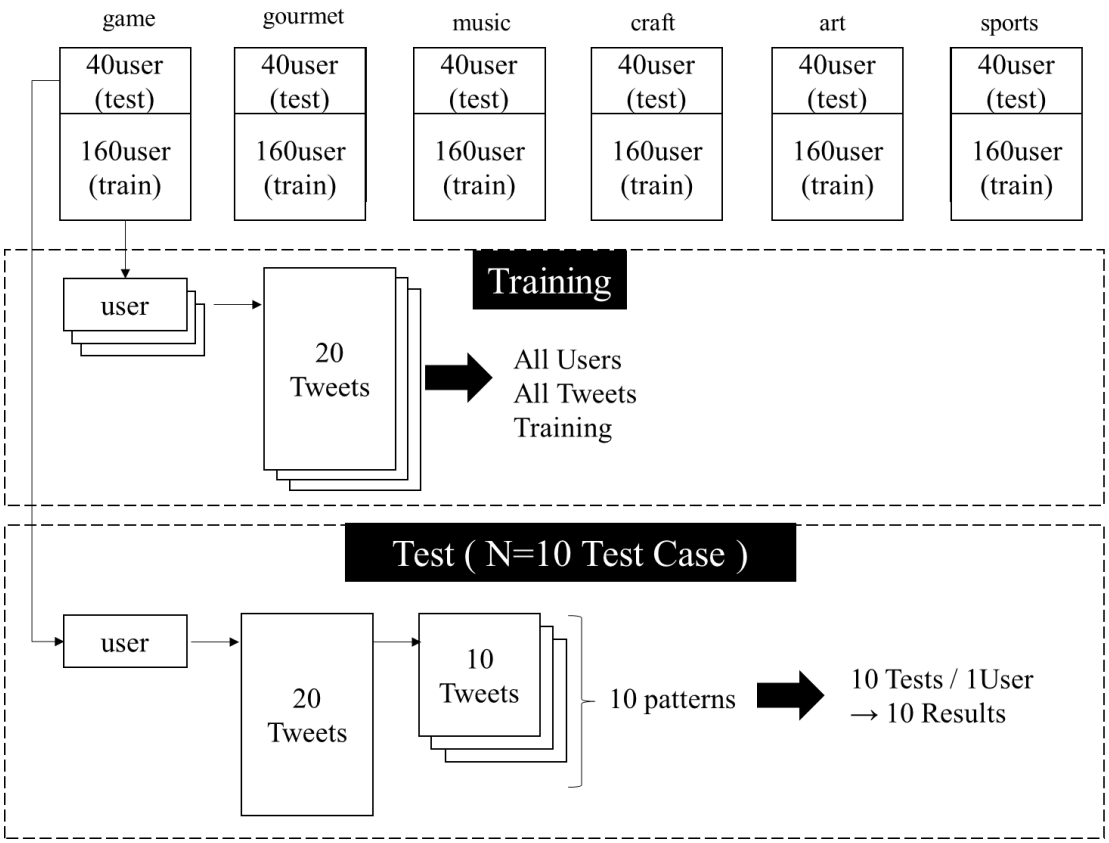


図 12 学習データと評価データの概要

4.2 ベースライン手法の実験結果

まずは Random Forests と SVM によるベースライン手法で測定した正確度について，特徴次元数ごとの結果を表 2 と図 13 に示す．Random Forests の結果では次元数 100 が最も高く，SVM の結果では次元数 70 が最も高いものとなった．平均値でも Random Forests の方が SVM よりも高い値となっており，総合して Random Forests の方が良い結果となった．

表 2 ベースライン手法の次元ごとの正確度

FeatureNo.	Random Forests	SVM
10	21.54	15.493
20	11.673	16.298
30	17.088	14.286
40	16.727	13.682
50	16.245	17.606
60	18.412	17.002
70	18.412	**20.926**
80	18.051	14.588
90	19.735	16.298
100	**22.623**	15.895
200	18.532	16.6
300	19.134	18.209
400	20.939	16.499
500	18.773	18.511
600	17.569	16.801
700	15.283	16.398
800	20.578	15.091
900	19.134	16.499
1000	19.374	17.304
2000	15.764	16.801
3000	15.283	17.404
4000	19.374	15.694
5000	16.847	16.197
6000	18.412	15.091
7000	17.81	17.505
8000	18.171	16.7
9000	20.578	16.197
10000	18.051	16.398
average	18.21828571	16.49903571

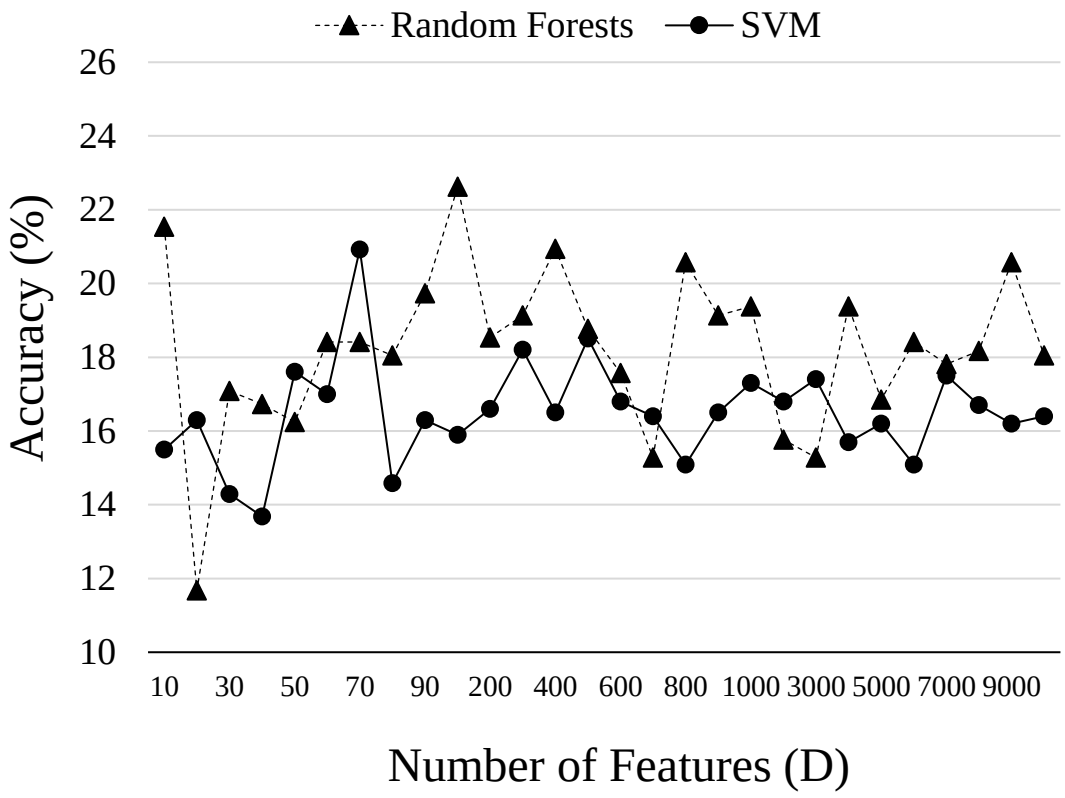


図 13 ベースライン手法の次元ごとの正確度

また, ベースライン手法での F1 スコアの結果も表 3 と図 14 に示す. Random Forests の結果では次元数 40 が最も高く, SVM の結果では次元数 50 が最も高いものとなった. 平均値でも Random Forests の方が SVM よりも高い値となっており, 総合して Random Forests の方が良い結果となった.

表 3 ベースライン手法の次元ごとの F1 スコア

FeatureNo.	Random Forests	SVM
10	18.711	14.786
20	21.204	13.267
30	22.221	19.726
40	**24.694**	17.602
50	14.98	**20.022**
60	18.896	17.312
70	12.644	16.765
80	17.175	18.132
90	16.328	16.289
100	16.298	17.45
200	15.129	18.459
300	16.816	19.589
400	16.108	14.11
500	16.235	15.979
600	18.434	15.376
700	14.189	16.824
800	18.188	14.944
900	19.842	13.656
1000	13.513	14.033
2000	17.182	15.219
3000	18.12	14.516
4000	17.82	15.752
5000	18.611	13.034
6000	17.452	13.685
7000	18.762	15.761
8000	14.267	15.092
9000	14.165	15.242
10000	18.783	12.904
average	17.38453571	15.91164286

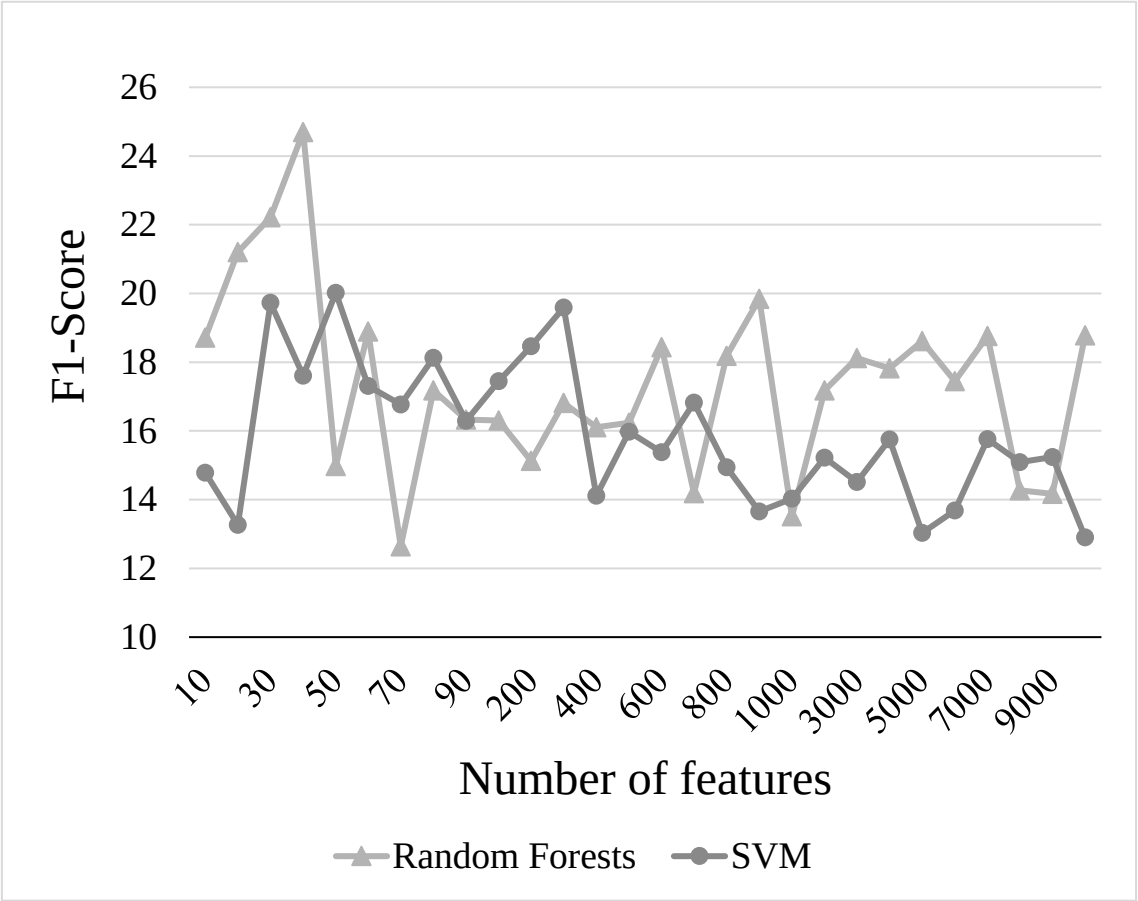


図 14 ベースライン手法の次元ごとの F1 スコア

4.3 提案手法の実験結果

使用するツイート数ごとの正確度について，提案手法の実験結果を表 4 と図 15 に示す.

表 4 ツイート数ごとの正確度測定結果

	Accuracy(%)				Number of Tweet
Number of Tweet	RNN	LSTM	GRU	CNN	average
3	32.1	39.185	40.376	24.765	34.1065
4	36.808	40.717	44.625	30.945	38.27375
5	35.493	42.173	44.332	31.039	38.25925
6	40.186	**46.347**	45.272	33.596	41.35025
7	37.413	43.067	45.391	37.258	40.78225
8	35.142	43.823	41.402	37.73	39.52425
9	36.455	41.69	43.159	37.649	39.73825
10	34.529	39.447	43.238	36.885	38.52475
11	38.084	43.473	42.754	38.802	40.77825
12	37.57	42.737	43.156	39.804	40.81675
13	39.698	40.034	42.044	38.526	40.0755
14	43.513	45.509	**45.709**	**41.517**	**44.062**
15	**44.359**	44.872	44.359	41.282	43.718
Method average	37.79615385	42.54415385	43.52438462	36.13830769	

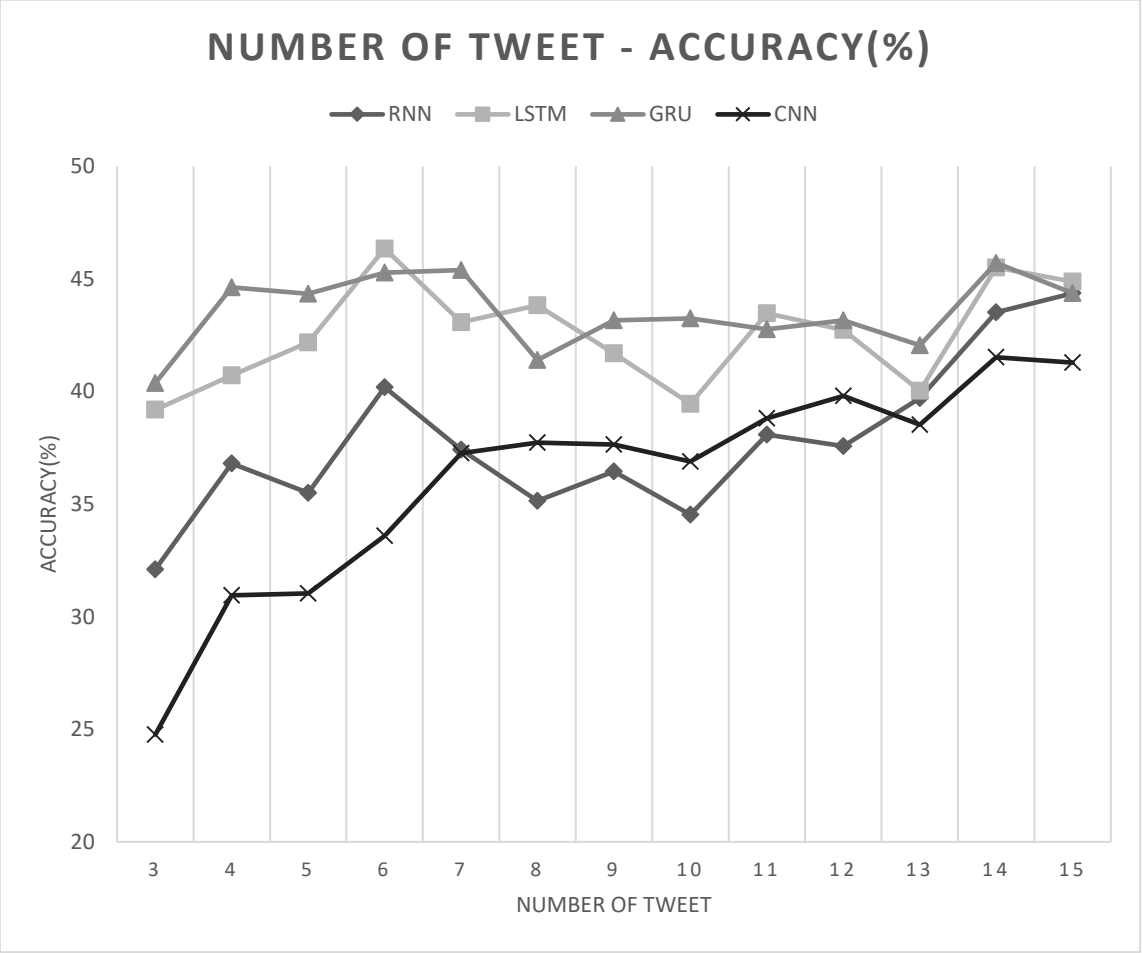


図 15 ツイート数ごとの正確度結果

この結果から，LSTM が最も高い正確度(46.35)を示していることがわかる．一方でベースライン手法(Bag of Words)は最も低い正確度を示した．しかし，GRU のほうが平均して高い正確度を示しており，LSTM と GRU にはそれほど差が見られなかった．RNN と CNN は学習ツイート数が増えると正確度が安定するが，ツイート数 N=6 未満までは，正確度が低い．

Random Forests によるベースライン手法では，特徴次元数 D が 100 のときに最も高い精度(22.628%)が得られたが，全体として 25%を下回り，分類の精度としては低いものとなった．LSTM(N=6)のときの学習曲線を図 16 に示す．

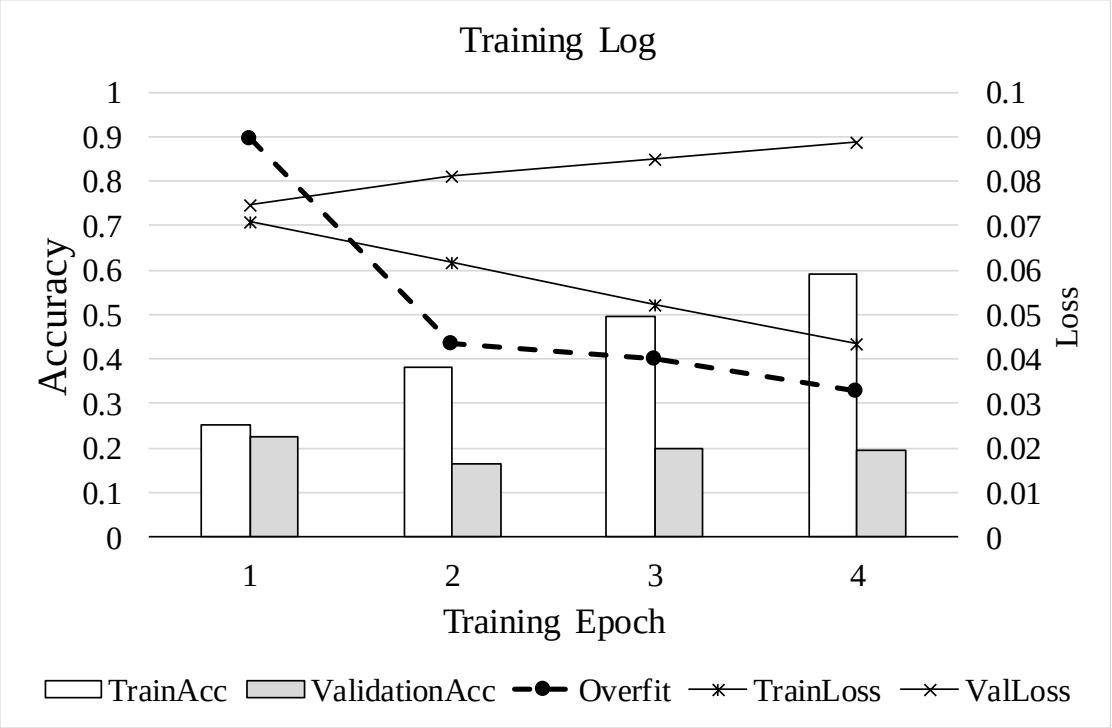


図 16 LSTM 学習結果（LSTM, N=6）

LSTM(N=6)のときのカテゴリ別の Precision, Recall, F1-score を表 5 に示す. この表からは, sports カテゴリの recall が高くなっていることが分かるが, その一方で, craft カテゴリは低い recall である.

表 5 カテゴリごとの結果 (LSTM, N=6)

	Precision	**Recall**	**F1-score**
game	39.30	48.65	43.48
gourmet	46.25	47.18	**46.71**
music	**56.84**	38.30	45.76
craft	49.59	38.46	43.32
art	53.29	51.74	52.51
sports	36.24	**65.41**	46.64

4.4 実験結果の考察

カテゴリ別の推定精度(Precision)をみると平均的に, music カテゴリが最も高い精度を示した. 実際のユーザのデータを見てみると, music に属するユーザの多くが, 何らかの楽曲作品の制作や演奏会をメインに活動をしていて, その活動の PR の場として Twitter を用いている傾向が顕著であった. 一方で, sports カテゴリには, 多岐にわたるスポーツが含まれており, スポーツを趣味でしているユーザだけでなく, スポーツ観戦の趣味なども含んでいる. スポーツ以外の趣味に関する発言が多く投稿されるなど, 共通する表現があまり含まれない傾向がみられた. この結果, Precision が最も低い結果となってしまったが, 様々なことを発言するため, 特徴量が分散し, Recall が高くなったと考えられる. このため, 投稿内容とプロフィールに関連の薄いユーザを学習データに含めてしまうことで, これがノイズとなり, 趣味カテゴリ推定が上手くいかないことも考えられる.

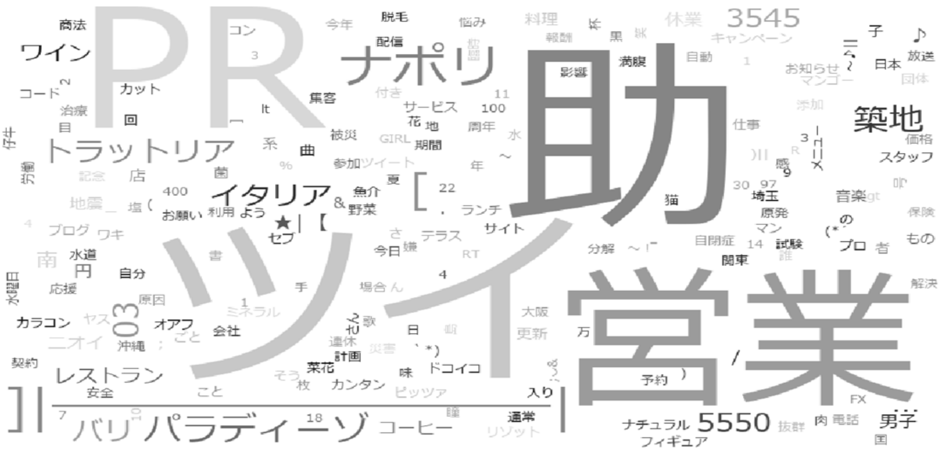
実際にニューラルネットワークの学習曲線を確認すると, Training Accuracy の最大値が約 60%となっており, 前述のようなノイズの除去を行わないかぎり, 学習データ量を増やしてもあまり改善が期待できないと考えられる. 図 17 に, χ^2値に基づく特徴選

択を使用して作成された各趣味カテゴリのワードクラウドを示す．素性の数は 100 に設定されている（名詞のみがワードクラウド上に表示されている）． 素性としてより特徴的な単語は，より大きなフォントで示されている．

図から見てとれるように，「仕事」という素性は game，craft，art のどのカテゴリにも頻繁に登場した．一方で，「アスリート」，「チャンピオンシップ」，「全日本」は sports のカテゴリにおいて特徴的な単語であり，「ナポリ」「レストラン」「ワイン」は gourmet のカテゴリにおいて特徴的な単語である． このことから sports と gourmet のカテゴリの精度は，他のカテゴリの精度よりも高くなったと考えられる．

game

gourmet



music

craft

art

sports

図 17 カテゴリごとのχ^2値から分析した Word Cloud

4.5 各カテゴリに分類したツイートとユーザの考察

最も正確度が高くなった LSTM(N=6)の混同行列を図 15 に示す. また, 表 4 から各手法のツイート数ごとの正確度を平均した場合, N=14 のケースが最も高くなる. このときの LSTM(N=14), RNN(N=14), GRU(N=14), CNN(N=14)のカテゴリ分類正答率を混同行列として図 18 から図 22 に示す.

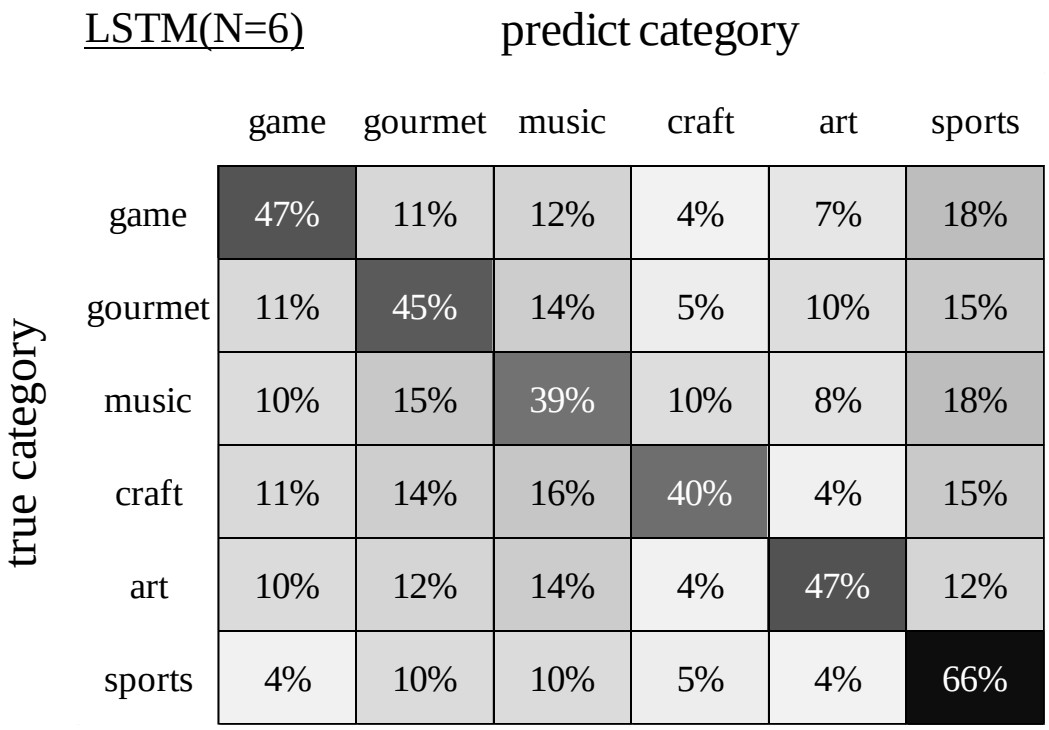


図 18 LSTM(N=6)の結果マトリクス

LSTM(N=14)

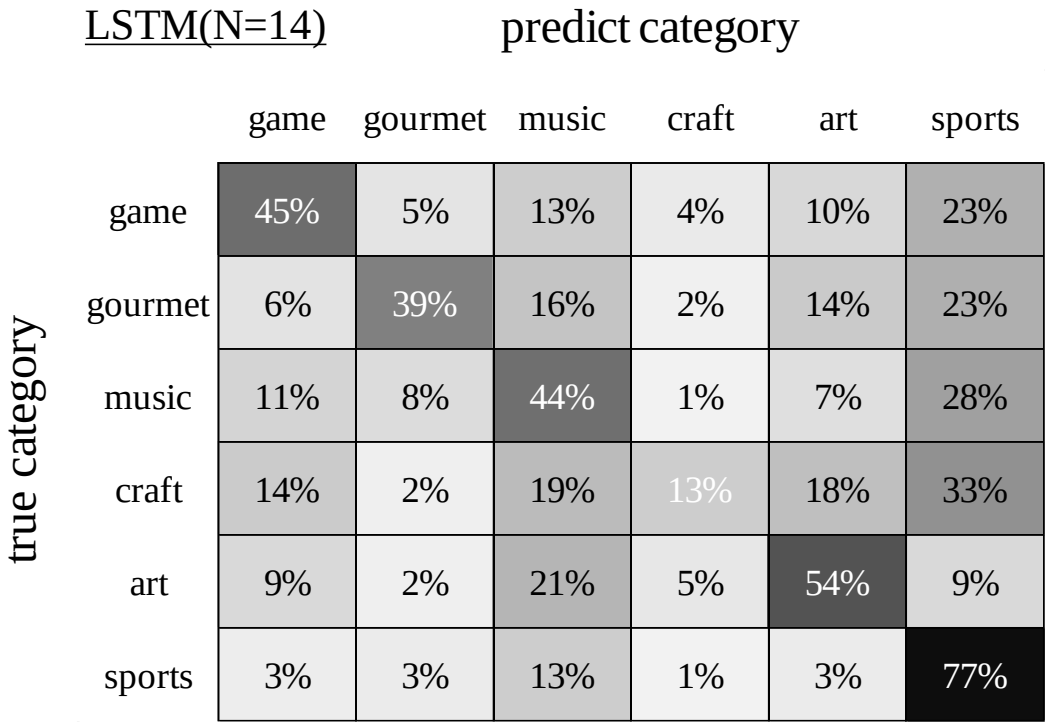


図 19 LSTM(N=14)の結果マトリクス

GRU(N=14)

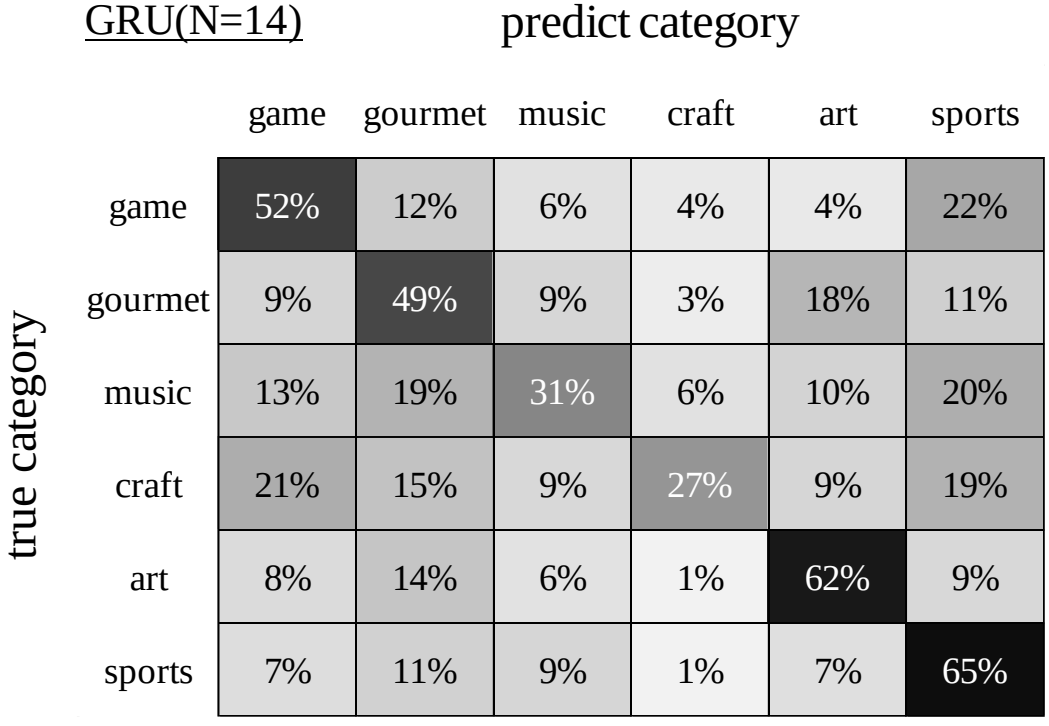


図 20 GRU(N=14)の結果マトリクス

RNN(N=14)

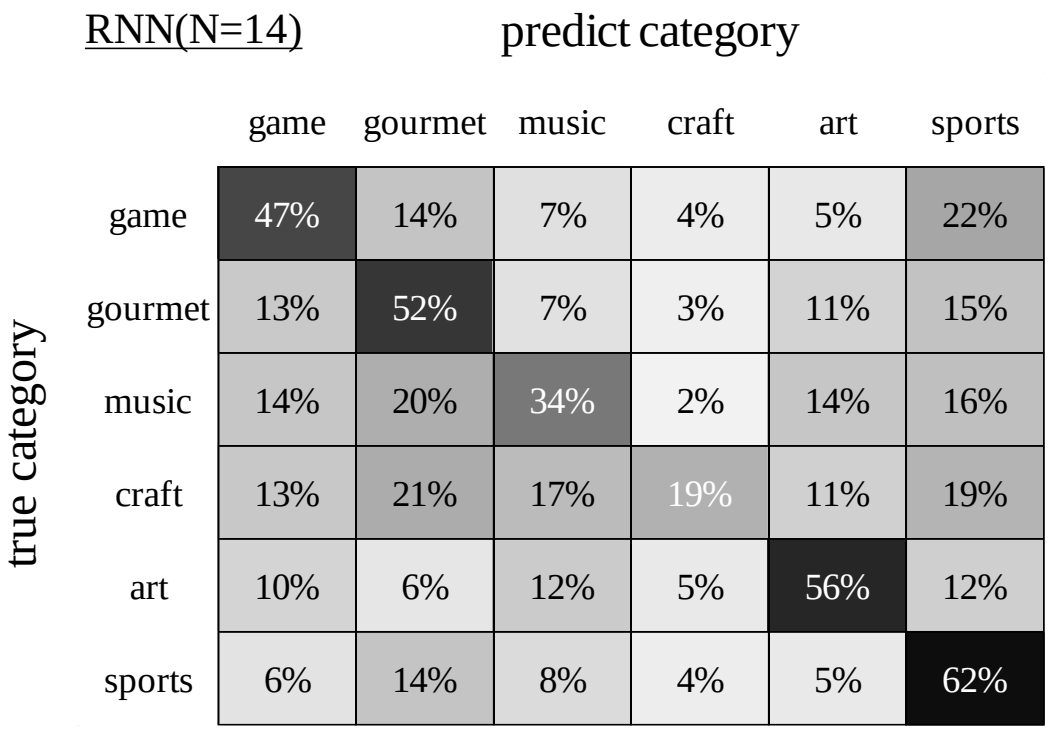


図 21 RNN(N=14)の結果マトリクス

CNN(N=14)

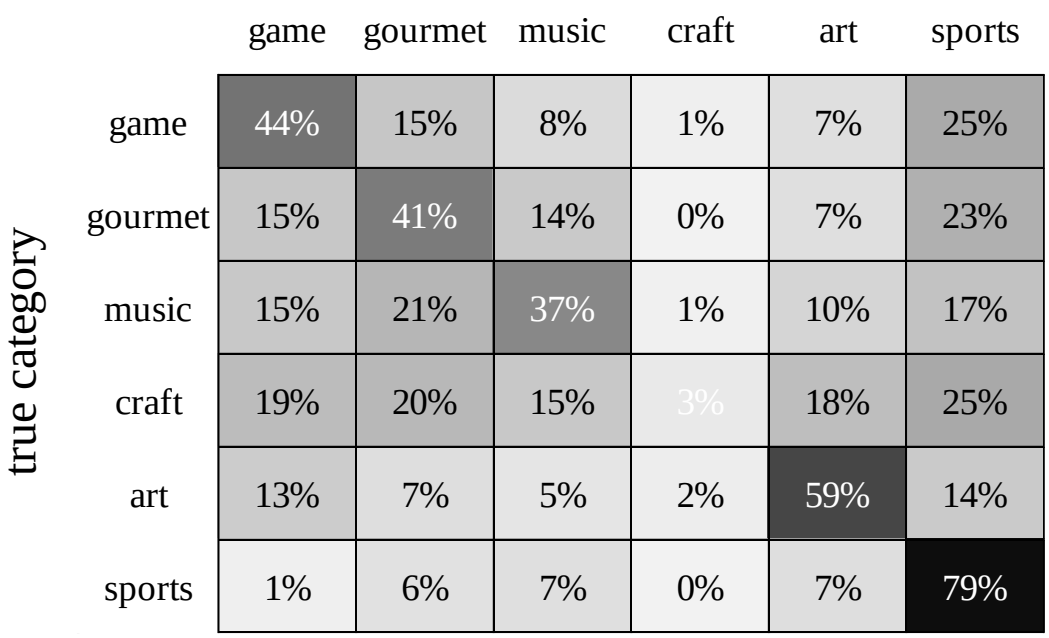


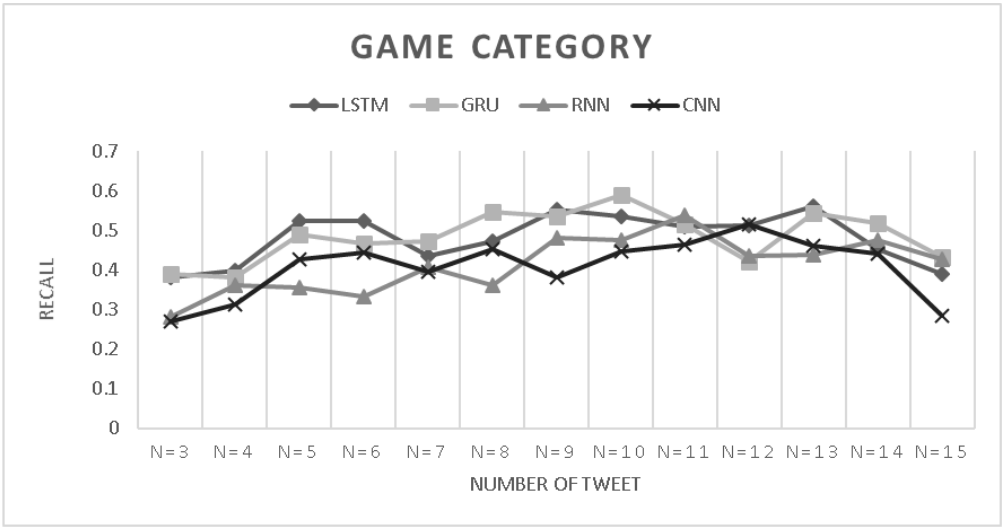
図 22 CNN(N=14)の結果マトリクス

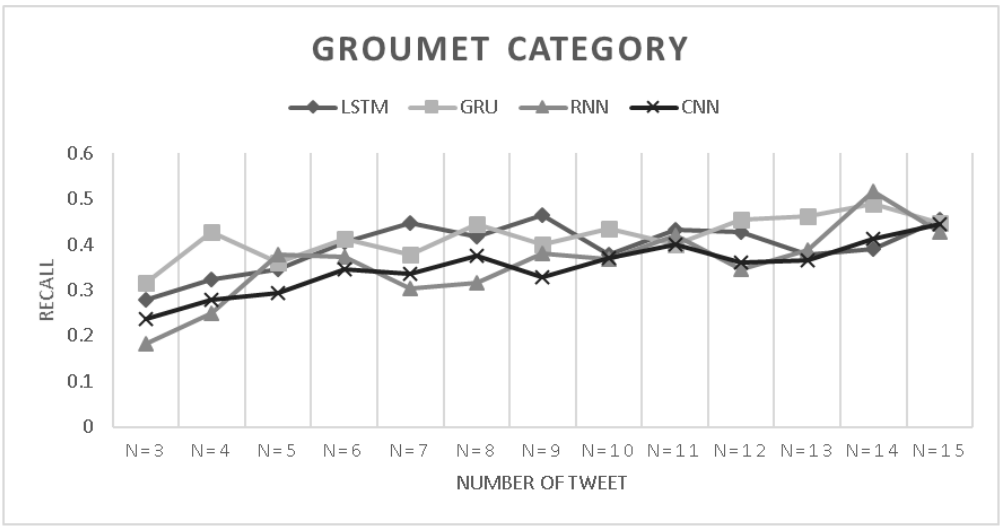
また，各手法のカテゴリごとの分類正答率を表 6 にまとめる．

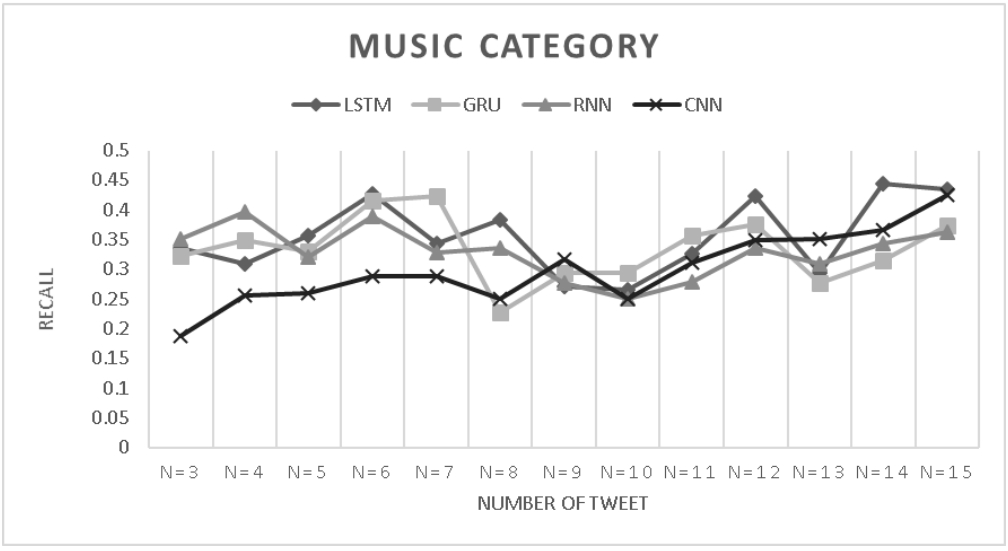
表 6 カテゴリごとの各手法の正答率

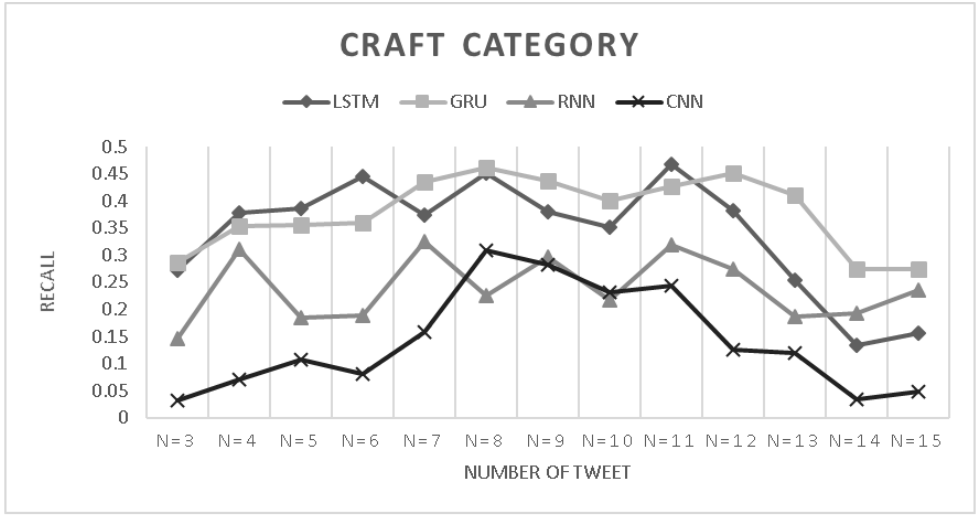
	LSTM	GRU	RNN	CNN
game	45%	52%	47%	44%
gourmet	39%	49%	52%	41%
music	44%	31%	34%	37%
craft	13%	27%	19%	3%
art	54%	62%	56%	59%
sports	77%	65%	62%	79%

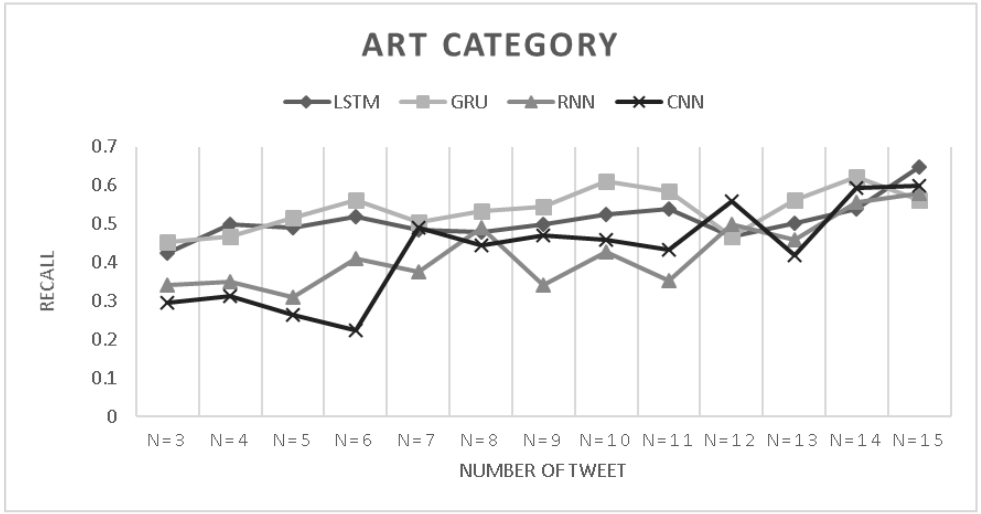
これら各手法の結果からみると，sports カテゴリの分類に関しては正答率高く分類ができている一方で，craft の正答率が著しく低くなってしまっている．LSTM(N=6)とLSTM(N=14)の結果から学習するツイート数が増えると craft の分類が難しくなっていくことが読み取れる．各カテゴリの分類正答率を手法ごとに図 23 にまとめる．詳細なデータは付録 A に示す．


GAME CATEGORY
LSTM
GRU
RNN
CNN
RECALL
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
N=3 N=4 N=5 N=6 N=7 N=8 N=9 N=10 N=11 N=12 N=13 N=14 N=15
NUMBER OF TWEET


GROUMET CATEGORY
LSTM
GRU
RNN
CNN
RECALL
0.6
0.5
0.4
0.3
0.2
0.1
0
N=3 N=4 N=5 N=6 N=7 N=8 N=9 N=10 N=11 N=12 N=13 N=14 N=15
NUMBER OF TWEET


MUSIC CATEGORY
LSTM
GRU
RNN
CNN
RECALL
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
N=3 N=4 N=5 N=6 N=7 N=8 N=9 N=10 N=11 N=12 N=13 N=14 N=15
NUMBER OF TWEET





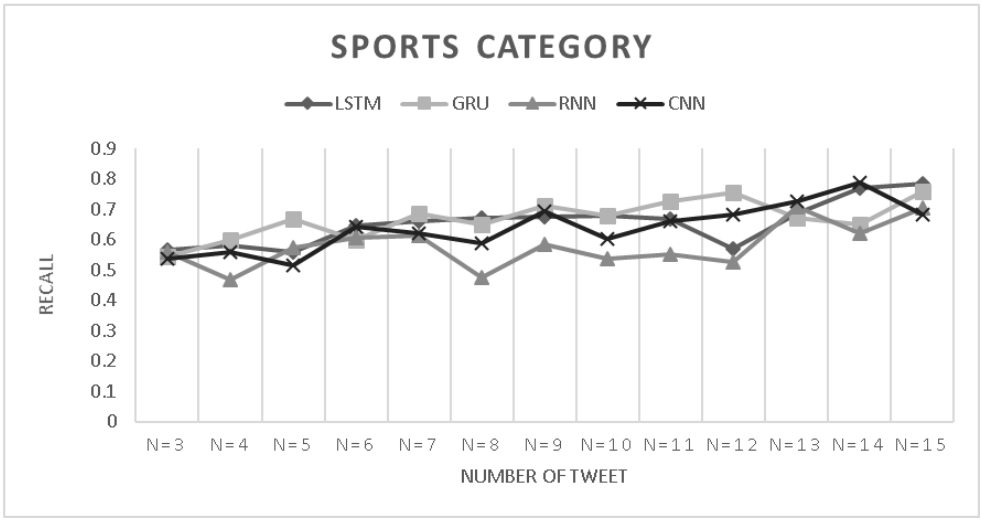


図 23 各手法の分類正答率グラフ

表 7 と図 23 の結果から，特に craft カテゴリは学習に使用するツイート数が増えると正答率が下がり，カテゴリ分類に失敗することが分かる．LSTM で craft カテゴリにおいて学習ツイート数が増えると正解から不正解に変わってしまうユーザをサンプリングして調査した．具体的なツイート内容は付録 B に示す．該当ユーザはロボット作りに関するツイートが多いが，ツイート数が増えると大会に関する情報も増えていき，試合についての情報が多くなった結果 sports カテゴリに誤って振り分けてしまったと考えられる．他には，人形作りの話題が詳細になるにつれ，作品としての情報が多くなり，art カテゴリに分類されてしまったと予想されるユーザの事例もあった．図 17 からみると「作品」というキーワードが craft と art カテゴリで大きなキーワードとなっていることが原因の一つではないかと考えられる．

また，craft カテゴリ自体に特徴的なキーワードが少なく，他のカテゴリにも興味を持つユーザが多かったのではないかとも予想できる．学習に使うツイートを増やすことにより情報量も増え，カテゴリ分類の正答率は向上すると直感的に予想できるが，カテゴリそのものの特徴が少ない場合，逆に分類が難しくなるケースがあることが分かった．craft カテゴリ分類に成功したユーザについても調査を行った．具体的なツイート内容は付録 C に示す．羊毛フェルトでの人形作りを趣味とするユーザで，いくつか craft に関連する特徴的なツイートがあることから分類に成功したと考えられる．該当ユーザについて詳細に調査したところ，学習ツイート数 N=3 のケースでは craft カテゴリ分類の正答率 33%であったが，N=5 から N=12 のケースでは 80%以上の正答率となっており，学習ツイート数が増えれば正答率が向上する成功パターンであることが分かった．

他のカテゴリについても学習ツイート数が増えると判定に失敗する例と，分類に成功する例について調査を実施した．sports カテゴリで見られた失敗例としては，スキューバダイビングを趣味とするユーザで冬の時期のツイートを抽出したためか，学習も評価も上手くいかなかった例が見られた．また，ツイートの期間も数週間単位で開いており，話題がバラバラである特徴もあった．このユーザは LSTM のみ分類に成功しており，残りの手法では間違ったカテゴリに分類されている．今回の実験では時系列順にツイートを抽出しているが，各々の時期については考慮していないため，季節柄に関する

趣味やツイートの時間的な間隔が開いている場合，話題の移り変わりなどで正しい連続性を得られないケースがあると考えられる．

また sports カテゴリにおいて判定に失敗しているケースは，学習・評価用のツイートにそもそもカテゴリに関係のないツイートばかりだった例が多かった．例となる具体的なツイートを付録 D に示す．このユーザについて調査をしたところ，プロフィールに「バレーボール」が表記されており，過去バレーボールの選手だったことが分かったため，広義的にスポーツに興味を持つことが正しいと判断できる．しかし直近のツイートに sports カテゴリに関するものが含まれていないため，ツイートからのカテゴリの分類に失敗してしまっていた．sports に関連する趣味があるとしても，観戦を趣味とする場合，試合がないタイミングやシーズンオフの場合は趣味推定に必要なツイートが取得できない可能性が高いと考えられる．このように，学習・評価するタイミングのツイート次第では判定が上手くいかないケースがあることがわかった．

次に sports カテゴリで分類が成功したユーザについて例を調査した．具体的なツイート内容は付録 E にあげる．該当ユーザはサッカーのスパイクについてレビューを紹介するもので，内容も関連するツイートで統一されているため分類が容易だったと思われる．このユーザのツイートは LSTM, GRU, RNN, CNN 全ての手法で sports カテゴリへの分類が成功しており，典型的な成功例となった．しかし，このように話題が統一されているユーザは営業用や広告用アカウントであるケースが多く，一般的なユーザは内容が分散していることもわかっている．多様な話題を持つユーザに対し，複数の連続ツイートから趣味を推定する方法は上手くいくケースもあるが，本項の考察で述べたように，推定の根拠とするツイート群を時期，間隔など適切に抽出することも重要な要因だといえる．

図 19 から LSTM での分類失敗は多くは sports カテゴリへの分類となっている．music カテゴリユーザを誤って sports に分類した例を調査した．具体的なツイート内容は付録 F に示す． 該当ユーザのツイート内容を見ると music に関連する話題が多く，sports に関する話題はない．このユーザは GRU では正しく分類できていたが，LSTM では失敗していた．ヒューリスティックには music に分類することができるが，ニューラルネッ

トワークによる学習結果では上記例のように分類失敗の原因を特定することが難しく，学習条件の調整や前提条件の見直しが必要になる．GRU の方が平均的には優れた分類精度となったが，LSTM の方が最高部類精度が高く，LSTM でしか分類できないものや，GRU でのみ分類できるものもあり，一概にどちらがよいという比較は難しい結果となった．

第 5 章　おわりに

本論文では, Twitter 上の連続した投稿内容を学習素性としてニューラルネットワークによりユーザの趣味カテゴリを推定する手法を提案した. RNN, LSTM, GRU, CNN の 4 種類のネットワークと, ベースライン手法として Bag of Words を特徴量とした Random Forest を用いた結果, LSTM を用いてツイート数 6 のときに最大の推定精度が得られた. 平均の結果では GRU の方が高い推定精度を記録したが, LSTM あるいは GRU のみでしか分類できなかったものもあり, 一概にどちらが優れた手法であるといえない結果となった.

また, 趣味カテゴリ gourmet が最も高い F1-score で推定できた. 一方で, 推定精度が 0%となるユーザもいた. この結果を分析したところ, 推定がまったくできないユーザのなかには, プロフィールに記載の趣味カテゴリとはまったく別の内容が連続して投稿されていた. これについては, 継続的にユーザのタイムラインを取得するか, プロフィール更新日に近いツイートを取得するなどすることで改善すると思われる.

今後の課題として, プロフィール内容とツイート内容の類似度に基づき, ツイートの取捨選択をすることによる精度改善を試し, 分散表現を既存のものから, 学習データに含まれるツイートで追加学習させることでより趣味カテゴリ推定に適した分散表現を作成することなどを予定している.

謝辞

本研究を行うにあたり，様々なご指導をご教授して下さった徳島大学大学院社会産業理工学研究部北研二教授に心から御礼申し上げます．また，本論文の作成について主査，副査をご担当いただき，ご指導を賜りました徳島大学大学院社会産業理工学研究部獅々堀正幹教授，徳島大学大学院社会産業理工学研究部泓田正雄教授に心から深く御礼申し上げます．

また，日々熱心にご教授，ご指導して頂いた徳島大学社会産業理工学研究部松本和幸准教授と徳島大学社会産業理工学研究部講師吉田稔講師に心から御礼申し上げます．

参考文献

[1] L. Sloan, J. Morgan, P. Burnap, and M. Williams, “Who Tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data,” PLOS ONE, Vol. 10, No. 3, pp. 1-20, 2015.

[2] F. Ren and K. Matsumoto, Emotion Analysis on Social Big Data, Vol. 15, No. S2, pp. 30-37, 2017.

[3] M. A. Magumba, P. Nabende, and E. Mwebaze, “Ontology boosted deep learning for disease name extraction from Twitter messages,” Journal of Big Data, Vol. 5, No. 31, pp. 1-9, 2018.

[4] M. D. A. Praveena and B. Bharathi, “A survey paper on big data analytics,” in Proc. International Conference on Information Communication and Embedded Systems (ICICES), 2017.

[5] M. K. Danthala, “Tweet analysis: Twitter data processing using Apache Hadoop,” International Journal of Core Engineering & Management (IJCEM), Vol. 1, issue 11, 2015.

[6] D. Sehgal and A. K. Agarwal, “Sentiment analysis of big data applications using Twitter data with the help of Hadoop framework,” in Proc. the International Conference System Modeling & Advancement in Research Trends (SMART), 2016.

[7] M. Kumar and A. Bala, “Analyzing Twitter sentiments through big data,” in Proc. the 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016.

[8] R. Vatrapu, R. R. Mukkamala, A. Hussain, and B. Flesch, “Social set analysis: a set theoretical approach to big data analytics,” IEEE Access, Vol. 4, pp. 2542-2571, 2016.

[9] S. Sohangir, D. Wang, A. Pomeranets, and T. M. Khoshgoftaar, “Big data: deep learning for financial sentiment analysis,” Journal of Big Data, vol. 5, no. 3, pp. 1-25, 2018.

[10] T. Forss, S. Liu, and K.-M. Bjork, “Extracting people’s hobby and interest information from social media content,” Terminology and Knowledge Engineering 2014, Berlin, Germany, 2014.

[11] M. Yu, X. Han, X. Gou, J. Yu, F. Lv, and J. Li, “Content-based social network user interest tag extraction,” International Journal of Database Theory and Application, vol. 8, no. 2, pp. 107-118, 2015.

[12] Y. Lewenberg, Y. Bachrach, and S. Volkova, “Using emotions to predict user interest areas in online social networks,” in Proc. IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 1-10, 2015.

[13] A. Agarwal, O. Rambow, and N. Bhardwaj, “Predicting interests of people on online social networks,” in Proc. International Conference on Computational Science and Engineering, Vancouver, BC, 2009, pp. 735-740.

[14] L. B. Krithika, “Finding user personal interests by Tweet-mining using advanced machine learning algorithm in R,” in Proc. IOP Conf. Series: Materials Science and Engineering, vol. 263, 2017, pp. 1-9.

[15] Makki, A. J. Soto, and S. Brooks, “Twitter message recommendation based on user interest profiles,” in Proc. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016.

[16] S. Volkova, Y. Bachrach, and B. V. Durme, “Mining user interests to predict perceived psycho-demographic traits on Twitter,” in Proc. IEEE Second International Conference on Big Data Computing Service and Applications (Big Data Service), 2016.International Journal of Machine Learning and Computing, Vol. 9, No. 2, April 2019.

[17] M. McPherson, L. Smith-Lovin and J. Cook,“Birds of Feather: Homophily in Social Networks,”Annual Review of Sociology, Vol. 27, pp.415?444, 2014.

[18] N. Mangal, S. Kanwar, and R. Niyogi, “Prediction of Twitter users’ interest based on Tweets,” in Proc. the First International Conference on Intelligent Computing and Communication, pp. 167-175, 2016.

[19] J. Li, A. Ritter, and E. Hovy, “Weakly supervised user profile extraction from Twitter,” in Proc. the 52nd Annual Meeting of the Association for Computational Linguistics, pp. 165-174, 2014.

[20] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth, “User interests identification on Twitter using a hierarchical knowledge base,” in Proc. the Semantic Web: Trends and Challenges: 11th International Conference, ESWC 2014, 2014, pp. 99-113.

[21] K. Watanabe and S. Kato, “Tweetrecommendation system reflecting user preference based on latent dirichlet allocation and collaborative filtering,” in Proc. the 28th Annual Conference of the Japanese Society for Artificial Intelligence, 2014, pp. 1-4.

[22] K. Kamijo, T. Natsukawa, and H. Kitamura, “Personality estimation from Japanese text,” in Proc. the Workshop on Computational Modeling of People’s Opinions, Personality, and Emotions in Social Media, 2016, pp. 101?109.

[23] K. Matsumoto, S. Tanaka, M. Yoshida, K. Kita, and F. Ren, “Ego-state estimation from short texts based on sentence distributed representation,” International Journal of Advanced Intelligence (IJAI), Vol. 9, No. 2, pp. 145-161, 2017.

[24] R. Kato, K. Nakamura, Y. Yamamoto, S. Tanaka, and K. Sakamoto, “Research for reasoning users’ attributes and habitual behavior of microblog,” IPSJ Journal, vol. 57, no. 5, pp. 1421-1435, 2016.

[25] K. Ikeda, G. Hattori, K. Matsumoto, C. Ono, and T. Higashino, “Demographic estimation of Twitter users for marketing analysis,” IPSJ Journal Consumer Device & System (CDS), vol. 2, no. 1, pp. 82-93, 2012.

[26] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, “Classifying latent user attributes in Twitter,” in Proc. the 2nd International Workshop on Search and Mining User-Generated Contents, 2010, pp. 37-44.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in Vector Space”.

[28] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in Proc. the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532?1543.

[29] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, “Bag of tricks for efficient text classification,” in Proc. the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2, 2017, pp. 427?431.

[30] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[31] K. Cho, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representation using RNN encoder-decoder for statistical machine translation,” in Proc. the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724?1734.

[32] Y. LeCunn, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436-444, 2015.

[33] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. the 3rd International Conference for Learning Representations, 2015.

[34] L. Breiman, “Random forests,” Machine Learning, vol. 45, issue 1, pp. 5-32, 2001.

[35] V. Vapnik, and A. Lerner, “Pattern recognition using generalized portrait method,” Automation and Remote Control, vol. 24, 1963. [44] Y. Yang, and J. Pedersen, “A comparative study on feature selection in text categorization,” in Proc. the Fourteenth International Conference on Machine Learning (ICML'97), 1997, pp. 412-420.

付録 A

提案手法（LSTM, GRU, RNN, CNN）ごとのカテゴリ分類正答率を学習ツイート数ごとにまとめた表を以下に示す.

game

Number of Tweet	LSTM	GRU	RNN	CNN
N=3	38%	39%	28%	27%
N=4	40%	38%	36%	31%
N=5	52%	49%	36%	43%
N=6	52%	47%	33%	44%
N=7	44%	47%	41%	39%
N=8	47%	55%	36%	45%
N=9	55%	53%	48%	38%
N=10	53%	59%	48%	45%
N=11	51%	52%	54%	46%
N=12	51%	42%	44%	52%
N=13	56%	54%	44%	46%
N=14	45%	52%	47%	44%
N=15	39%	43%	43%	28%

groumet

Number of Tweet	LSTM	GRU	RNN	CNN
N=3	28%	32%	18%	24%
N=4	32%	43%	25%	28%
N=5	35%	36%	38%	29%
N=6	41%	41%	37%	35%
N=7	45%	38%	30%	34%
N=8	42%	44%	32%	37%
N=9	47%	40%	38%	33%
N=10	38%	44%	37%	37%
N=11	43%	40%	42%	40%
N=12	43%	45%	35%	36%
N=13	38%	46%	39%	37%
N=14	39%	49%	52%	41%
N=15	45%	45%	43%	44%

music

Number of Tweet	method			
	LSTM	GRU	RNN	CNN
N=3	34%	32%	35%	19%
N=4	31%	35%	40%	26%
N=5	36%	33%	32%	26%
N=6	43%	42%	39%	29%
N=7	34%	42%	33%	29%
N=8	38%	23%	34%	25%
N=9	27%	29%	28%	32%
N=10	27%	29%	25%	25%
N=11	33%	36%	28%	31%
N=12	42%	38%	34%	35%
N=13	30%	28%	31%	35%
N=14	44%	31%	34%	37%
N=15	44%	37%	36%	43%

craft

Number of Tweet	method			
	LSTM	GRU	RNN	CNN
N=3	27%	29%	15%	3%
N=4	38%	35%	31%	7%
N=5	39%	36%	18%	11%
N=6	45%	36%	19%	8%
N=7	37%	43%	32%	16%
N=8	45%	46%	23%	31%
N=9	38%	44%	30%	28%
N=10	35%	40%	22%	23%
N=11	47%	43%	32%	24%
N=12	38%	45%	27%	13%
N=13	25%	41%	19%	12%
N=14	13%	27%	19%	3%
N=15	16%	27%	24%	5%

art

Number of Tweet	method			
	LSTM	GRU	RNN	CNN
N=3	43%	45%	34%	30%
N=4	50%	47%	35%	31%
N=5	49%	52%	31%	26%
N=6	52%	56%	41%	22%
N=7	48%	51%	37%	49%
N=8	48%	53%	49%	44%
N=9	50%	54%	34%	47%
N=10	52%	61%	43%	46%
N=11	54%	58%	35%	43%
N=12	47%	47%	50%	56%
N=13	50%	56%	46%	42%
N=14	54%	62%	56%	59%
N=15	65%	56%	58%	60%

sports

Number of Tweet	method			
	LSTM	GRU	RNN	CNN
N=3	57%	54%	56%	54%
N=4	58%	60%	47%	56%
N=5	56%	67%	57%	51%
N=6	65%	60%	61%	64%
N=7	66%	69%	61%	62%
N=8	67%	65%	48%	59%
N=9	68%	71%	59%	69%
N=10	68%	68%	54%	60%
N=11	67%	73%	55%	66%
N=12	57%	75%	53%	68%
N=13	69%	67%	71%	73%
N=14	77%	65%	62%	79%
N=15	79%	76%	71%	68%

付録 B

LSTM（N=14）によるカテゴリ分類において、正解カテゴリ craft への分類に失敗してしまったツイートを以下の表に示す.

ツイート日時	ツイート内容
2017/12/18 11:05	バトンタッチを受けました. 新しい中の人です(^_-)-☆||これからは県内各地で開催されるデモなどの情報を発信したりしていこうと思います. ||これからも沖縄高専ロボット製作委員会のことをよろしくお願いします!!
2017/12/22 23:15	おはようございます. |今日は全ロボと東北交流会がありますね!|沖縄高専からは全ロボに 3 名,東北交流会に 1 名が参加します. |参加者には頑張ってもらい, しっかり技術や知識をアウトプットできるよう願っています. |ぜひ, 交流頑張って下さい… https://t.co/FEGgJoYibo
2017/12/22 23:45	一方で, 今日は沖縄高専のある名護市の隣に位置する宜野座村の宜野座村総合体育館にてデモが行われます!|デモでは地区大会にてしなる竿で防御をしていた, 沖縄 A のクジャクロボなどを披露します!お時間のある方はどうぞお越し下さい!(^∀^)ﾉ https://t.co/jcKv1HFPKx
2017/12/23 00:50	ちなみに, 沖縄 A の宝物は 2 千円札を持ったキジムナーでした!|キジムナーはガジュマルという木の精霊と沖縄では言われています. |デザイン力の高い部員が製作しました(^^ゞ https://t.co/kocROUZPFM
2018/01/18 12:08	明けましておめでとうございました(今更)|新年初投稿ですね. |明日から 2 日間, 豊見城市民体育館で開催される「IT 津梁まつり 2018」の土曜日の方で沖縄高専のヤンバルクイナ型ロボを出展致します!|ｺﾝﾊﾞﾝﾜ！(*ﾟ▽ﾟ)ﾉﾞ☆ﾊﾟﾁﾊﾟ… https://t.co/BuDeEenBJg
2018/01/27 03:27	9 年前(2009 年)の九州沖縄地区大会の会場にて現在行われている, 名護サイエンスフェスタ 2018 にてデモ中です. https://t.co/iPYp5LLINj

2018/02/10 06:13	浦添市民体育館にて青少年科学作品展が開かれています．その中で沖縄高専の高専ロボコン出場ロボのデモ中です．|2 日間の日程で開催しているため，明日も開催しています!もし良ければご覧ください!||※ただいま，期末試験期間中のため人が居な… https://t.co/SSzI5LOBuw
2018/02/11 08:13	この 2 日間，子どもたちにたくさん操縦されたクジャクもお疲れです．|おかげで，多くの人たちに高専のことやロボコンを知ってもらえたと思えば嬉しいです!!||ちなみに，クジャクの羽(竿)ですが地区大会前の練習で中の人が 1 度折っただけでそれ以外で折れたことがありません!|タフな羽でした!
2018/02/15 11:56	本日は 5 年生の追い出し会をしました!|多くの部員がお世話になった先輩，中の人は迷惑をかけたり基礎を教えて頂いたり深い思い出のある先輩たちでした!|その分，悲しいですね……(ノД`)ｼｸｼｸ||また，本日は試験明けで疲れていますが，… https://t.co/GCrg4OsymW
2018/02/23 12:12	鹿児島空港へ着きました．|明日から都城高専にてプレ MJ ロボコンがあります．|前日である今日は鹿児島空港に降り立ち，今から都城市へ向かいます!||明日は都城高専さん，よろしくお願いします!!||飛行機の離陸時に中の人は，はしゃいでし… https://t.co/8XGxxJOawF
2018/02/25 13:50	MJ ロボコン終了しました(*’∀’)|午前の部はマシンが上手く動かず結果が出せませんでしたが，||午後の部では，1,2,3 年生 4 人で構成したチームが準優勝，1 年生だけのチームがベスト 4 入りと，かなり良い結果でした!||それ以外にも… https://t.co/EqJ7WhqhPO
2018/02/25 13:59	何よりも都城高専の人たちが優しくて工具を貸していただきマシンをお見せしてくれたり，移動手段の無い僕らの送迎をしてくれたり，優しくしていただきました．|地区大会の時もエアーコンプレッサを貸していただきましたし，もっと仲良くしたいです… https://t.co/dQMQiAsx3L
2018/03/01 07:24	本日は名護小学校にてデモを行っていました!||予想以上に多くの児童生徒が集まり，ロボットで風船を割ってみたい人を募集すると「はー

	い!!!, はーい!!!」という大声がたくさん聞こえて懐かしき小学生の元気を感じました. ||2 枚目の写真… https://t.co/CyKsCWdnp2
2018/03/17 12:33	今日は卒業式でした|卒業する先輩たちと話し, 退職される顧問の先生にエールを頂き, 嬉しくも悲しいです. |機会があればまた顔を出してほしいですね. ||さて, 3 月 22 日はイオン名護店にて都城高専とのミニロボコン,交流を行います. |お時間… https://t.co/UuKeoPAAZq
2018/03/22 11:42	ミニロボットフェスティバル, 大盛況でした!|都城高専からは 2 名の学生に来ていただき, ミニロボコンをしました. |相変わらず都城高専の「のばせ!ピンポンパンダ」が強かったです. 決勝再試合も制し, 2 連覇!|悔しい思いもありますが, この思… https://t.co/KOILAs9Rgj
2018/04/13 12:56	新年度初ツイートですね(・∀・)||4 月 16 日(月)の 17 時から夢工場 2 階にて新入生向けの説明会をします!!||部活動紹介でも強調したように, 学科なんて気にしないで下さい!|ロボコンが気になってる, 興味がある方はどうぞお越し下さい… https://t.co/LBWdXBeRCn
2018/04/14 08:54	みなさん周知の通り, 学内でインフルエンザが大流行しています. それに伴う休校により, 新入生向けの説明会も遅らせていただきます. |後ほど, 日時が決まり次第にご連絡します m(_ _)m||それにしても補講などが増えて活動に支障が出る恐れ… https://t.co/pxY445PGcC
2018/04/22 14:42	報告が遅くなってしまいましたが, 明日の 23 日 17 時から夢工場 2 階に変更します!!||ぜひ, お越し下さい!!||1 年生に届いてくれーー(> ω <) https://t.co/PVYJmwtCb1
2018/04/26 11:17	ついに来ましたね……!!|ルール発表です!!||普段とは大きく違うイメージのルールですが, 今年もどんな試合の様子が見られるのか非常に注目ですね(≧▽≦)||私たちもまずは国技館を目指して, ロボット製作頑張っていきます!!|応援よ… https://t.co/Qok7u3xK8H
2018/05/20 11:39	遅くなりましたが, 本日は新入生歓迎会を行いました!||試験も近いですが, 1 年生には今回の歓迎会でパワーを付けてこれから頑張ってほしい

	ですね(≧▽≦)‖毎年行っているBBQをしたのですが，中の人は皆さんに飯テロできるような写真を撮… https://t.co/jA8tS3LLc0

付録 C

LSTM（N=14）によるカテゴリ分類において、正解カテゴリ craft への分類に成功したツイートを以下の表に示す.

ツイート日時	ツイート内容
2016/05/27 16:07	今日はカオスな天満へ. |ﾒはクーロンズバー. ネットで見つけて, ゲームのクーロンズゲート好きの為, 行ってみた. |*|オーナーさんもクーロンズゲートが好きで, しかもセカンドライフ(仮想世界)でクーロンズバーをやっていて, それをリアルに… https://t.co/FrklnyaPah
2016/06/20 08:42	いつも, ひんやりシートにおてての肉球だけぺたりとつけてる. なぜ？（笑）|.|#エキゾチックショートヘア #エキゾチック #エキゾ #exoticshorthair #exotic #tw https://t.co/WPnLdye4BC
2016/06/25 06:46	久しぶりに欲しい物が‥. 試飲したい. |DeLonghi(デロンギ) 全自動エスプレッソマシン マグニフィカ S（ ECAM23120B ） https://t.co/QDt5UwJIaF @さんから
2016/07/02 12:38	RT @XXXX: 今年2月ベルリン国際映画祭で賞を取った映画にリアル猫ヘッドが出ています. Grüße aus Fukushima | Trailer 1 | Deutsch HD German (Doris Dörrie) https://t.co/oqgZn2…
2016/07/02 12:38	RT @XXXX: 桃井かおり出演の東日本大震災後の福島を舞台にしたドイツ映画『フクシマ・モナムール』(ドリス・デリエ監督）が「ハイナー・カーロウ賞」を受賞!!! 第 66 回ベルリン国際映画祭|#桃井かおり #ニュースクリップ #ベルリン国際映画祭| https:/…
2016/07/02 12:38	RT @XXXX: «In Japan bin ich immer noch ein Elefant» https://t.co/w67uTYlTuE

2016/07/10 12:11	参議院選投票してきました. 与党が過半数か…😱|こうなると, もう私には政治も国民も殆どが理解できない(´°　ω°　)チーン　|#tw https://t.co/AL0ODt5bPK
2016/08/01 13:21	RT @XXXX: 7月26日(火)から NHK スタジオパーク／「岩合光昭の世界ネコ歩き番組展」で私の制作した巨大リアル猫人形が展示されています. 猫好きの方はこの機会に横幅2メートル超の巨大なリアル猫人形と一緒に記念写真を撮りましょう. https://t.co/q…
2016/08/01 13:23	RT @XXXX: Tw の皆さんリアル猫ヘッド販売情報の拡散大感謝です. こちらも苦労して制作したものなのでお忘れなく！NHK スタジオパーク／「岩合光昭の世界ネコ歩き番組展」に巨大猫！出現！|https://t.co/6E1I0pKf2y https://t.co/…
2016/09/05 07:42	RT @XXXX: 羊毛毡里的萌猫世界(羊毛フェルトでつくる ウチのコそっくりかわいい子猫)|中国語に翻訳された書籍が届きましたˆˆ中国ではすでに販売されているようです. https://t.co/oCSbHClsrq https://t.co/ah3YApPv8u
2016/09/05 07:42	RT @XXXX: 10月からヨミカル教室が5つになります. 柏, 北千住, 大森, 恵比寿, 自由が丘, 全ての教室で私が指導します. お近くの猫好き犬好き動物スキの皆さんぜひこの機会にご参加ください. https://t.co/cp8UMKrZhQ https://t.co/…
2016/09/23 04:32	私の友人の娘さんです. クラウドファンディング残り 8 日です！ご協力お願いします！|アルペン・植野琴／祖父と恩師へ捧げるオリンピック, 私が勝たなければならない使命！	ACT NOW https://t.co/y79SXibeNa
2016/09/24 12:15	RT @XXXX: 10月からヨミカル北千住で「羊毛フェルトで作るリアル犬猫教室」が始まります. 第2週 月曜日 12:30〜15:30 10/10, 11/14, 12/12, 1/9, 2/13, 3/13 生徒募集中です. https://t.co/2kSUDWdMqg…

2016/10/07 13:30	コーヒージャンキーなので，全自動エスプレッソマシン，デロンギ マグニフィカ S ECAM23120B 買いました．美味しくて楽チン！… https://t.co/PdO1czIB9I
2016/10/07 13:53	RT @XXXX: DJ みそしると MC ごはんが食器洗いをラップ！ あのリアル猫ヘッドも登場 #kai_you https://t.co/5cNPTQ1moi
2016/10/07 13:53	RT @XXXX: 10 月 17 日から読売カルチャー大森でウチのコそっくり羊毛猫人形教室教室が始まります．生徒募集中です．https://t.co/rcL4572erH https://t.co/3tt4q0divL
2016/10/23 10:35	ドールズミス行ってから京橋のジュニームーンへ．最近発売されたブライスを見に…☺|結局ヘンリエッタ買いました💦|アウトフィットが全然趣味じゃなくてスルーしていたけど，髪の毛の色とふわふわ加減が気に… https://t.co/Ll0I6HLWBl
2016/12/01 12:42	RT @XXXX: いつのまにか海外のまとめサイトで一万件以上シェアされてた．https://t.co/ZsoM0zvvCB @XXXXX さんから
2016/12/01 12:43	RT @XXXX: ネコ・・・|이달의소녀탐구 #28 (LOO ΠΔ TV #28) https://t.co/036YT8Fzje @YouTube さんから
2017/09/13 12:17	えっ！チャッピーさん？！一番好きな日本のジョンでした‥ご冥福をお祈りいたします． https://t.co/5qC7m4xu4A

付録 D

LSTM（N=14）によるカテゴリ分類において、正解カテゴリ sports への分類に失敗してしまったツイートを以下の表に示す.

ツイート日時	ツイート内容
2018/05/17 15:24	バブリーにゃんコーヒー☕|福岡！飲みに！観に来てね！ https://t.co/dGMOCxdQSi
2018/05/17 15:25	今日のやっ P～！ https://t.co/caGtKOqGk0
2018/05/18 08:53	衝撃おったまげ写真がぶっ飛びーーー！！瞬き禁止だゾ🤡🌬 https://t.co/d87hxqUTFo
2018/05/20 12:17	おったまげーーー！！！|くりそつゲロマブ女発見！|ご一緒にリッツパーティいかが？ https://t.co/5PZsT6C061
2018/05/21 09:56	👺🌟👺🌟👺🌟 https://t.co/BCY0AlsnIt
2018/05/23 04:49	庶民のみなさんごきげんよう|今日のやっ P～ですぅ～|あ～たたちぃ～いいね押してごらんなさ～い https://t.co/cTMdll5ldu
2018/05/24 11:13	今日のやっ P～！ https://t.co/PlCSIY5my4
2018/05/24 22:56	5 月 26 日（土）放送|21:00～フジ TV||『さんま&女芸人お泊り会～初めて後輩に語る, 62 年走り続けた男の人生哲学～』||おったまげーな番組！！|しょう油顔とゲロマブ女芸人たちをお見逃しなく！|あたしもちょっとだけ出演してるゾ！ https://t.co/qsrKCl1TCY
2018/05/25 15:00	今日のやっ P～！|(㊗ゆりヤンまつ毛エクステ) https://t.co/FmUasO3ZDR
2018/05/30 09:24	RT @XXXX: 2017.12.31🖥ミラモン～✨ ⑩|平野ノラ🕺の「スランプ脱出特訓開始‼」|# 石川真佑 |# 平野ノラ https://t.co/0jJ3PnQxkD

2018/05/30 09:24	RT @XXXX: 2017.12.31🖥ミラモン～✨ (11)|特訓終了‼ノラ, 真佑ちゃんの凄さ認める ٩("ˬ)9✧ |#石川真佑|#平野ノラ https://t.co/FRcdlx1Pv3
2018/05/31 15:47	この後 25:49～日テレ||「採用！フリップ NEWS」||観てね🌥☺ https://t.co/T46CWBGGVs
2018/06/01 11:08	今日のやっ P～！ https://t.co/2PwxhpK0ZZ
2018/06/02 00:34	今日のスペシャルやっ P～！|(上から富士山 JP) https://t.co/pkM5HjOWcC
2018/06/02 00:35	平野ノラ出演||🌥2 日「さんまのお笑い向上委員会」23:10～フジ||🌥4 日「あさイチ」8:15～NHK||🌥5 日「ヒルナンデス」11:55~日本テレビ||🌥5 日「採用！フリップ NEWS」23:59～中京テレビ||🌥7 日「採用！フ… https://t.co/XhfyfpLWxu
2018/06/02 14:42	バブリー速報！|たけしさんからオファシャルグッズの T シャツ頂いちゃったゾ❤|ありがとうございます！|要チェケバブリーーー！！！||「KITANO BLUE」||https://t.co/tACjPg2qJd||#KITANOB… https://t.co/0acwYyXrYx
2018/06/04 01:28	今日のやっ P～！|(市川 右團次さんと歌舞伎ポーズ) https://t.co/EOhcfm6mjm
2018/06/05 11:33	今日のやっ P～！ https://t.co/RFaUwPp5R9
2018/06/09 14:30	平野ノラ出演||🌥9 日「さんまのお笑い向上委員会」23:10～フジ|🌥10 日「こんな田舎がアルか否か！？」10:00~テレ朝|🌥12 日「ヒルナンデス」11:55~日テレ|🌥14 日「採用！フリップ NEWS」24:54～読売テレビ… https://t.co/qpUbjO29oO
2018/06/12 04:25	♫トラトラノラ～ https://t.co/GVRiPINZdA

付録 E

LSTM（N=14）によるカテゴリ分類において、正解カテゴリ sports への分類に成功したツイートを以下の表に示す.

ツイート日時	ツイート内容
2018/06/01 07:19	NGナイキ サッカースパイクNG|『マーキュリアル スーパーフライ VI FG』|ナイジェリア代表カラー公開！|2018W 杯でナイジェリア代表の選手が着用|https://t.co/XnT2krIxDy https://t.co/yR2GnXUcDs
2018/06/01 08:58	㊍PUMA 新型 サッカースパイク ㊍|『プーマワン 1 IL レザー』&|『フューチャー2.1 NETFIT』|2018 ワールドカップモデル公開！|6 月 1 日 17 時から発売開始|https://t.co/WlBDMstsOw https://t.co/oyqLLv7v7Q
2018/06/02 11:24	アディダス サッカースパイク|『エックス 18.1 ジャパン HG/AG』|機能性・素材・特長・考察記事|https://t.co/7ym4l1ry4p https://t.co/qRFmgrrDkw
2018/06/04 12:26	1994 年～2014 年まで|ワールドカップに合わせて公開された|歴代ナイキ CM を振り返ってみた！|https://t.co/AS0x2L99Tx|(↓動画は前回の Winner Stays On) https://t.co/yvZIjwDSEd
2018/06/05 08:35	ニューバランス サッカースパイク|『MiUK ONE FG Paul Smith』|ポールスミスとの限定コラボモデル|MADE IN UK (イングランド製)|海外では 2018 年 6 月 5 日発売… https://t.co/88djtO5dcx
2018/06/05 11:54	ナイキ サッカースパイク|『JUST DO IT PACK』|ヒールの国旗カスタマイズについて考察|https://t.co/ScH93naF76 https://t.co/68rdEOfven
2018/06/06 11:03	【アディダス ゴールデンブーツ】|2018 ロシアワールドカップの|得点王に贈られるトロフィーは|adidas『エックス 18+』がモチーフ✨|https://t.co/Pz0LspUB8v https://t.co/xnAjSbSwwf
2018/06/06 11:58	ミズノ サッカースパイク|海外版『モレリア』|白×赤の新色が海外で発売！|https://t.co/0ZodLEZ96p https://t.co/7ZAhGhqg4L

2018/06/07 11:13	ミズノ サッカースパイク|『モレリア 2 & モレリア IN』|『モレリア NEO2』|レッド×ホワイト(赤×白)カラーを考察！|https://t.co/Bx9qwByKaq https://t.co/igKZUL6sGq
2018/06/07 21:30	NIKEiD サッカースパイク|『マーキュリアル スーパーフライ 6』|『マーキュリアルヴェイパー12』|国旗追加 & カラバリ UP して新登場！|https://t.co/NNI3ZMZW1l https://t.co/BZfgxN6kqa
2018/06/08 11:53	ナイキ サッカースパイク|『ティエンポ レジェンド 7 エリート HG』|機能性・素材・特長・考察記事|https://t.co/cUdfSK1txT https://t.co/GoSxhdKRdu
2018/06/09 11:34	ミズノ サッカースパイク|レビュラ 2 V1 JAPAN (27,000 円)|レビュラ 2V1 (￥21,600 円)|細かな違いや選び分けのポイントを徹底比較レビュー！|https://t.co/ZnG9jXmAme https://t.co/2pk8LMICXY
2018/06/10 11:47	ナイキ サッカースパイク|『ハイパーヴェノムファントム 3 HG』|サイズ感や新 HG ソールなどを考察！|https://t.co/CpGEVRqxD8 https://t.co/nllLp6OMCH
2018/06/11 09:38	アルゼンチン代表 FW ディバラ|2018 ロシア W 杯での着用スパイクは|『アディダス GLITCH 2.0』|https://t.co/70WkWxMlHU https://t.co/WwAsFn5wBB
2018/06/12 12:08	スイス代表 MF グラニト・ジャカ|2018W 杯での着用スパイクは|『アンダーアーマー マグネティコ PRO』|ジャカ専用 シグネチャーモデル！|https://t.co/rXEnAhIbkk https://t.co/I51wG096vv
2018/06/13 11:08	ナイキ サッカースパイク|『マジスタ オーパス 2 エリート FG』|サイズ感や FG ソールなどを考察！|https://t.co/rioIcUy5kA https://t.co/EYvApqyTka
2018/06/14 09:40	アディダス サッカースパイク|『プレデター18+ FG/AG GR』|2018 ロシア W 杯開催記念 限定カラー|Gosha Rubchinskiy コラボモデル|https://t.co/MYtyt2yEkl https://t.co/muzHnHIg6z

2018/06/15 10:24	RT @XXXXX: 君はいくつ覚えている?スパイクマイスターKohei が選ぶ歴代スパイク #マイベストイレブン https://t.co/VdGcIGceQX #gekisaka https://t.co/AeaVqeqowJ
2018/06/15 11:06	ベルギー代表 ロメル・ルカク|PUMA と正式契約を締結！|『プーマワン 1 IL レザー』を着用|https://t.co/Z4YI5tuLEA https://t.co/Yv3rHvuNiq
2018/06/15 20:09	ポルトガル代表 C・ロナウド|スペイン戦でハットトリック！|着用スパイクはナイキ|『マーキュリアル スーパーフライ VI エリート FG』|#POR #ESP #W杯 https://t.co/35VbjrnPzo

付録 F

LSTM（N=14）によるカテゴリ分類において、正解カテゴリ music への分類に失敗してしまったツイートを以下の表に示す.

ツイート日時	ツイート内容
2018/05/08 01:57	RT @XXXX: 今アナログレコードの人気が注目されています. CD が人間に聞える約 22kHz より高音をカットしているのに対し, ハイレゾ音源などを超高音域まで再生できるため, 脳が心地よさを感じます. なお昨年米国で発売された LP 版『交響組曲 AKIRA』も即高評…
2018/05/08 14:25	僭越ながらエゴサしたところ,「ハイパーソニック・エフェクト」はまだまだ一般的ではないようです. |体験したらわかる／気付くその存在, 誤解なく広まればと一組員ながら願っています.（「組員」の語感, とてもやべぇ）… https://t.co/dWSjAaxHU4
2018/05/08 14:27	リンクが切れますね, , , . |このハイパーソニック・エフェクトの発見者が, 科学者としての大橋力先生こと山城祥二先生であるんであるんであります.
2018/05/08 14:39	誤解というのは大げさですが一例を申し上げますと, 通常のイヤホン・ヘッドホンでは高周波音は出ず, デジタルに出力する設備にはかなりの手間と金額が必要です. |また, 単音では効果もなく, 耳だけから聞くことも良くない, 等のこともありますので, 詳細は上記著作にて, 科学的な実験を御覧願います.
2018/05/08 14:52	長文失礼しました. |そして, では生音の体験は, ということで, 今週日曜は説明会, ケチャ稽古もありますのでぜひお気軽にお越しください！|#ケチャ #ガムラン #合唱 #民族音楽 #芸能山城組||https://t.co/OU0o6NyJTY
2018/05/15 00:02	RT @XXXX:【雑誌掲載 INFO】|「夏 Walker 首都圏版 2018」にケチャまつりが掲載されました！| 今年は 8/1(水)〜8/5(日)新宿三井ビルディング

	55HIROBA で開催，入場無料です．新田真剣佑さんが表紙の夏 Walker は首都圏の書店，コンビ… merged_timelines.txt: 3046669327 996178850223505408 2018-05-15 00:02:00 RT @fashionpressnet: [明日開催] 渋谷パルコの工事仮囲いに『AKIRA』の新アートワーク展示 - https://t.co/5skzusooKl https://t.co/f2hL7VKHPt
2018/05/15 01:08	RT @studentysg: 誤解というのは大げさですが一例を申し上げますと，通常のイヤホン・ヘッドホンでは高周波音は出ず，デジタルに出力する設備にはかなりの手間と金額が必要です．|また，単音では効果もなく，耳だけから聞くことも良くない，等のこともありますので，詳細は上記著作…
2018/05/17 15:52	RT @XXXX:【WEB 掲載 INFO】|Time Out Tokyo さんに『芸能山城組ケチャまつり』を取り上げていただきました！新宿三井ビル 55 HIROBA で無料で見られるケチャまつり．43回目の今年は8/1(水)〜8/5(日)の開催です．|https:…
2018/05/17 15:52	RT @TimeOutTokyoJP:【New!】ケチャは新宿で見る．『芸能山城組ケチャまつり』は 2018 年 8 月 1 日 - 2018 年 8 月 5 日まで 新宿三井ビル 55HIROBA にて開催．https://t.co/I1EO61Sb3I https://t.co/IigYUlrJ…
2018/05/19 14:20	RT @XXXX: 渋谷パルコの工事仮囲いを利用したアートウォール．こちらでは第1弾と第2弾の両方を動画でアップしてくださっています．大友克洋の「AKIRA」と河村康輔の共同作品が展示されています．https://t.co/0kwWfZV22G
2018/05/22 02:56	RT @XXXX: いま注目の『交響組曲 AKIRA』のハイレゾ版 LP！|芸能山城組の説明会と体験稽古では，その驚異の音源が生まれた秘密の一端を知ることができます！近くは 5 月 27 日，6 月 10 日，6 月 24 日ほか↓ https://t.co/fg9623O0vU|#…
2018/05/22 16:31	いやぁ，なかなかツイッターを見られなくなっている今日この頃の中の人です．恐縮です．|さて睡眠．最近は寝付きも悪うございます．森の中で，生き物たちの音の中に埋もれて，寝たいものですね（急に詩的）

2018/05/22 16:32	このアカウントにも特性というか，キャラ付けをしたいですね．どうしましょう，今後旧仮名づかいで呟くとか…．
2018/05/24 15:17	「雷神不動北山櫻」，千穐楽も近いですね．幕見でしか拝見できていないのですが，諧謔味も，演者観客相互の阿吽の呼吸もバツグン‼|その演目に着想し，山城先生などが生み出した，誰でも参加できるのに迫力・阿吽の呼吸では本物に勝るかとも思え… https://t.co/mkbdirDRqJ
2018/05/27 16:00	ツイッターも開けないようでは精神不衛生でございます．テスト的に投稿してみますが，下記はケチャ稽古の今後の日程です！！|逆に，あなたはなぜケチャをやらないんですか？ https://t.co/yr0LAZyGbt
2018/06/01 16:02	RT @tora0820: 初めてのデートで思わず AKIRA 用語が出る女． https://t.co/QVa0kUewb5
2018/06/07 02:59	RT @XXXX:【稽古日記】|リハーサルでは組頭から合唱やガムラン演奏，楽器の配置，PA まで入念な指示があり見違えるように！タンザニア・ゴゴ人の合唱曲では，組頭の「もっと不埒な感じで」のひと言で，おおらかな大地の歌声が湧き上がってきました．本番では最高のも…
2018/06/07 02:59	RT @XXXX:【稽古日記】| 3 日は，6 月 20 日に高校生に向けての公演のリハーサルでした．搬入や設営は全員で力を合わせて行います．バリ島の旗や城門等を設置すれば，気分もどんどん高まってきます．本番での舞台の転換も大事なパフォーマンスなので，しっかりと段取…
2018/06/13 03:50	RT @XXXX:【ネットラジオ】|芸能山城組の組頭・山城祥二が，英ネットラジオ局 NTS　Radio の特番として大友克洋氏『AKIRA』の音楽に影響を与えた曲の選曲を担当しました．|現在アーカイブとトラックリストが公開されていますので，ぜひお聞きください．…