5V-04

BERTを用いたマイクロブログユーザの興味推定に関する研究

中川　嵩将†　　上田　芳弘†　　坂本　一磨†
公立小松大学　生産システム科学部†

１．はじめに

現代社会において，ソーシャルメディアの普及に伴い，一般の人々の情報発信が活発となり，ネット上には大量のデータが存在している．一方で，投稿者の性別や年齢，職業といった属性を明示しない場合が多いため，この情報を有効的に活用するために属性推定の研究が盛んに行われている．属性の中でも興味は，ユーザの需要を大きく反映し，マーケティングに活用する上で非常に重要な要素だと考えられる．具体的には，音楽に興味があるとするなら，聴くための周辺機器やCDが必要となる．本稿では，ソーシャルメディア上のテキストデータから興味に関する属性推定を行う．これまでに興味に近い属性の趣味嗜好を推定する研究[1]では，Web 日本語 N グラムコーパスの嗜好カテゴリと入力文章の所有物表現の共起頻度を求め，所有物ごとに最も頻度が高い嗜好を関連付けているが，単語の頻度で解析しているため，文脈理解ができていない．本稿で使用する BERT は 2018 年に Google 社から発表された[2]ニューラル言語モデルであり，さまざまな言語タスクで最先端のモデルを大きく上回る性能を示している. BERT は，学習時にあるトークンを処理する際，他の全てのトークンの情報を参照して処理される．同じトークンでも，文章よって含まれるトークンが異なるため，文脈に対応した処理がされる．この Transformer による双方向の学習することで，文脈をより考慮することを可能としている．上記の特徴を持つ BERT を用いて，投稿者の興味推定に活用可能なテキスト分類手法を開発する．

２．研究概要

本提案手法の概要を図１に示す．本システムは，テキストデータの整理機能，テキスト分類用のファインチューニング，テキストデータのカテゴリ推定処理で構成される．

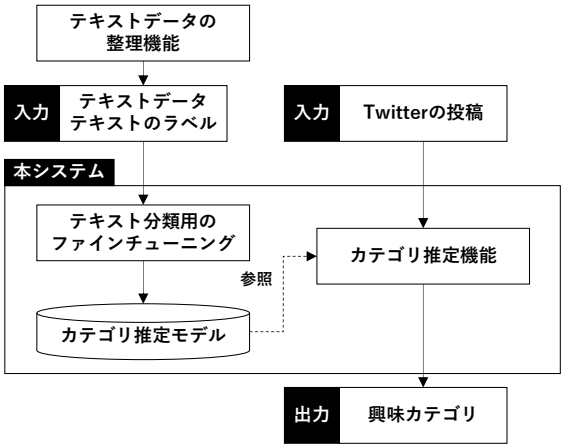

図１　本システムの概要

２．１　テキストデータの整理機能

本処理では，ファインチューニングで入力するためのラベル付きテキストデータを作成する．
まず，人の興味をカテゴリ化するには，大量のカテゴリが必要となることが考えられるが，本稿ではネットニュースのカテゴリに絞って検討する．また，テキストデータもネットニュースから Web スクレイピングし，下記の構成で保存する．

・１行目：記事のタイトル
・２行目：記事の URL
・３行目：記事の本文

２．２　テキスト分類用のファインチューニング

本処理では，テキストデータの整理機能で取得したテキストデータを BERT に学習させ，分類モデルを構築する．BERT には事前学習とファインチューニングがあり，事前学習では大量のデータが必要なものの，本稿で使用する BERT は，日本語の Wikipedia 記事を学習した事前学習済みモデル[3]を使用する．ファインチューニングでは，比較的少数のラベル付きデータを用いて，本稿ではテキスト分類に特化するように学習する．

２．３　カテゴリ推定機能

本処理では，構築したテキスト分類モデルに Twitter の投稿を分類させ，投稿者の興味を推定する．

Research on Estimation Interest of Microblogging Users Using BERT
† Takanobu Nakagawa , Yoshihiro Ueda and Kazuma Sakamoto
Faculty of Production Systems Engineering and Sciences,Komatsu University

３．実験概要

本実験では，Twitter の投稿を BERT に入力し，各投稿をカテゴリ分けし投稿者の興味を推定する．

３．１　実験内容

本実験で使用するネットニュースは， 2021 年 12 月 15 日～12 月 21 日に収集した Yahoo!ニュースを使用する．また，カテゴリは国内，国際，経済，エンタメ，スポーツ，地域の 6 カテゴリとし，記事数は 479 件である．このテキストデータをファインチューニングでは，6 割を学習データ，3 割を検証データ，1 割をテストデータとして使用する．Twitter の投稿は，10 人の投稿からそれぞれ 100 件ずつ無作為に抽出したものを使用する．

３．２　結果と考察

ファインチューニングによるテストデータの分類精度は，約 8 割の精度で推定されることが分かった．学習時のデータに対する損失の値を図 2 に示す．縦軸に損失の値，横軸に処理したミニバッチの数を示す．また，薄い線が実際の出力，濃い線が tensorboard のスムージングの出力を示す．学習データの損失が高いのは，学習データに使用する記事数が足りないことが考えられる．

実験結果は，表 1 には興味推定の結果，表 2 に分類結果の上位 2 つまでと人手による分類を示す．人手による分類の評価は，人がそのテキストを確認し，上位 2 つ程度のカテゴリを判定する．また，Twitter の投稿の分類では，全ての投稿者においてエンタメが最も多い結果となった．出力結果の中でもスポーツの分類は，他のカテゴリと比べ精度が高いと感じた．ニュースの中でもエンタメとスポーツは特徴があるため精度が高くなったと考えられる．表 2 では，推定結果と人手による分類が 1 つのカテゴリが合致している場合，正解だと判断すると．約 9 割で正解している結果となる．一方で，エンタメには日常会話や食事，音楽など本実験にないカテゴリが含まれるため，カテゴリを増やす必要がある．

４．おわりに

本研究では，Twitter の投稿から BERT を用いた興味推定を行った．本実験では，学習データの損失が高く，表 1 の評価も未だ不十分であるため，学習データやカテゴリを増やし興味推定の評価の検討を進めていく．

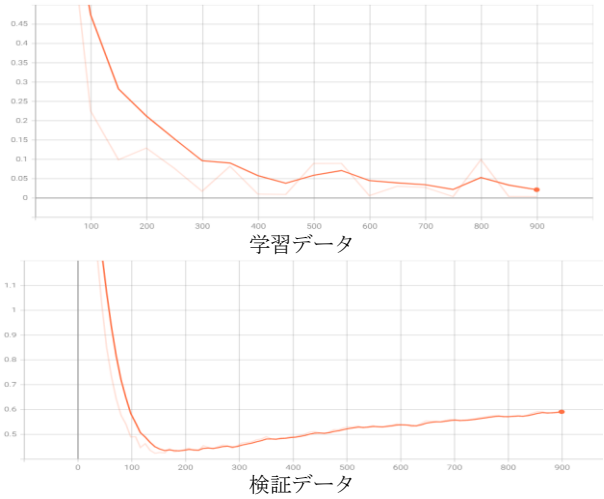

図 2　学習時のデータに対する損失

表 1　興味推定の結果

		カテゴリ					
		国内	国際	経済	エンタメ	スポーツ	地域
投稿者	A	4	2	4	60	0	30
	B	33	1	10	40	0	16
	C	12	1	3	80	0	4
	D	13	2	3	64	1	17
	E	15	2	12	62	3	6
	F	16	10	1	66	0	7
	G	6	1	3	66	0	24
	H	13	0	4	58	0	25
	I	24	1	19	26	2	28
	J	15	1	5	47	1	31

表 2　興味分類

		カテゴリ	
		推定結果	人
投稿者	A	エンタメ・地域	エンタメ・地域
	B	エンタメ・国内	エンタメ・経済
	C	エンタメ・国内	エンタメ・国内
	D	エンタメ・地域	エンタメ・地域
	E	エンタメ・国内	エンタメ・スポーツ
	F	エンタメ・国内	エンタメ・国内
	G	エンタメ・地域	エンタメ・国内
	H	エンタメ・地域	エンタメ・地域
	I	エンタメ・地域	国内・地域
	J	エンタメ・地域	エンタメ・国内

参考文献

[1] 馬縁，徳久，寺嶌.ユーザの嗜好と所有物の関係性を用いた属性分析.情報処理学会研究報告. Vol.2014,IFAT-114. No.7.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanava..BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.NACACL-HIT,2019

[3] “Pretrained Japanese BERT models released / 日本語 BERT モデル公開” https://www.nlp.ecei.tohoku.ac.jp/news-release/3284/