ツイートテキストデータによるリツイート数予測とその要因分析

増川哲太[1] 雨宮正弥[1] 仲田明良[1] 高須遼[2] 狩野芳伸[1,2]
[1]静岡大学 情報学部 [2]静岡大学 総合科学技術研究科情報学専攻
[1,2]{tmasukawa,mamemiya,anakada,rtakasu,kano}@kanolab.net

概要

SNS 投稿が社会に及ぼす影響は、人々の行動を左右するまでに大きくなった．Twitter に投稿されたツイートの影響を測る指標の一つはリツイート数である．本研究では投稿テキストのみからリツイート数の予測を試みた．結果、リツイート数が十分に多い場合、フォロワー数の貢献は低くテキストのみで予測可能であることを示したうえ、我々が大規模ツイートデータで事前学習した RoBERTa が他のモデルより良く、対数正規化したリツイート数に対し絶対平均パーセント誤差で 0.602 の予測性能であった．ツイート文中の各文節の予測への貢献度合いを分析し、「バズる」ツイートの傾向について示唆を得た．

1 はじめに

現代社会において、ソーシャルネットワーキングサービス(Social Networking Service、以下 SNS)は我々の生活に深く根付いており、経済から政治まで様々な分野に世界的な影響を及ぼしている．行政機関から企業、一個人に至るまでの多様な情報の発信において、SNS を使用することは今やスタンダードなものとなっている．そのような SNS の中でも、Twitter[i]は活発にテキストの投稿が行われていると同時に、データ取得 API が公開されており、膨大なデータを収集分析することができる．

Twitter ではリツイート機能により対象のツイート（投稿）をフォロワー全員と共有できるため、リツイートは情報の拡散において重要な要素である．先行研究の多くはフォロワー数、フォロー数、いいね数や文長といった数値的特徴のみを用いてリツイート数の予測を行っているが、投稿内容を無視した予測は不正確な可能性が高い．ツイート本文の言語的特徴からリツイート数を予測し、どのような文章が拡散されやすいかの分析ができれば、予測性能向上のみならずツイートがどのような社会的影響をもちうるかの分析が可能になる．

そこで本研究では、投稿テキストの特徴量を用いたリツイート数の予測を行った．予測には我々が大規模ツイートデータで事前学習した JTweetRoBERTa をファインチューニングして使用しほかの事前学習モデルと比較したところ、最も良い性能を得た．また、フォロワー数の影響も比較した結果、リツイート数が十分に多い場合、投稿テキストのみほうがよい性能を得ることができた．さらに文節の Ablation Study により、リツイート予測値を決定づける要因を分析した．

2 関連研究

Sharma ら(2022)[1]は、人間のリツイート行動パターンを把握するために、文字数や単語数などツイートの本文の数値的特徴と、フォロワー数やフォロー数などのユーザープロフィールの数値的特徴からリツイート数の予測を行った．これら数値的特徴量を組み合わせた新たな特徴量を提案して、リツイートの予測を行った．

Liu ら(2014)[2]は中国の Twitter に似た SNS である Weibo[ii]について、同様のユーザープロフィールや投稿に関する情報を用いてリツイート数を予測した．訓練データをリツイート数でいくつかのクラスに分割した後、テキストとユーザーに関する数値的特徴量とクラスラベルを用い、クラス分類タスクとして学習、推測した．

こうした数値的特徴を用いてリツイート数の予測を行う手法では、ユーザーの思考や興味を反映させた予測できているとはいえない．Twitter ユーザーの思考がもっとも強く映し出されているのはツイートテキストそのものであり、それを用いることでよりリツイート数をより正確に予測できると期待する．

Zhang ら(2016)[3]はユーザーのツイートを類似し

[i] https://twitter.com/

[ii] https://m.weibo.cn/

た単語やトピックでクラスタリングし、それらの注目度と現在のツイートの類似度やツイートの文脈情報、その他特徴量を学習させ、ある特定のツイートをリツイートする可能性を予測している. Wang ら(2020)[4]は Weibo から大量のデータを収集し、ユーザーの投稿本文、プロフィール、ユーザー間のネットワーク情報のベクトルをつなぎ合わせて分類することで、ある投稿をリツイートするかどうかの予測を行っている．これらの先行研究では、Twitter ユーザーの興味・関心を言語特徴やその他特徴を用いて取り込んでいるが、ユーザー個別のリツイート行動を予測することが目的であり、本研究の目的であるリツイート数の予測とは異なる.

3 提案手法とデータセット

本研究では、我々が独自に 6 億件の日本語ツイートで RoBERTa-base を事前学習した **JTweetRoBERTa**[5]の利用を提案手法として、東北大学が公開している Wikipedia で事前学習した日本語 BERT[6]の **BERT-base**[iii]、同 **BERT-large**[iv]、rinna 社の提供する Wikipedia および cc100 で事前学習した **RoBERTa**[7]の三つを比較対象として用いた.

これら事前学習済みモデルに対して、以下で説明する Twitter データセットによるファインチューニングを行った．具体的には、BERT の最終層に回帰分析用の全結合層を追加してファインチューニングを行った．ファインチューニングの各種パラメータ設定は、付録 A1 に記載する.

「インフルエンサー」であること自体がリツイート数に影響を及ぼす可能性を分析するために、対象ツイートを発信したアカウントのフォロワー数のみでの予測と、**JTweetRoBERTa** の最終全結合層にフォロワー数も入力した場合の予測を行った.

さらに、ツイートのどの文節がリツイート予測に重要かの Ablation Study による要因分析を行った.

3.1 ファインチューンの変数と損失関数

ファインチューニング時の正解となるリツイート数 r に幅があり学習が難しくなってしまうため、$x = log_2(r + 1)$ と正規化した値 x を用いた．以下 x を対数正規化リツイート数と呼ぶ.

リツイート数のスケールの幅が大きいため、誤差を比率で表現する損失関数も検討したが、正解値と予測値によっては簡単に小さい値になりやすくモデルの最適化には向かない、そのためモデル学習時の損失関数としては平均絶対誤差を採用した.

3.2 Twitter データセット

本研究で使用したデータは、Twitter API v2vの Academic Research アクセスにより取得した．ツイート本文がリツイート数にどのくらい影響を及ぼすのかをテキストに焦点をおいて分析するため、リプライ、引用リツイート、画像・動画を含むツイートは除外した．URL や絵文字は削除した.

リツイート数の少ない投稿ほど頻度が高いが、対象投稿のリツイート数が十分に多くないと統計的な予測が難しい．特にリツイートのない投稿はその理由に特殊な要因が考えられるため、すべて除外した.

ファインチューニング用データセット　日本語ひらがな五十音を含むツイートを、制限を設けずタイムライン上から収集するランダムサンプリングで、投稿期間は 2022 年 1 月 1 日~2022 年 10 月 31 日である．しかしこの方法のみでは、リツイート数が多いツイートを十分に取得することが困難だった．そのため、リツイート数が多かった投稿をまとめてさらにリツイートしているアカウントをいくつか収集し、そこでリツイートされた投稿をすべて収集したもの（以下「まとめアカウントデータ」と呼ぶ）を加え、リツイート数の多い投稿を補強したものをファインチューニングに用いた．ただし、ひらがなないしカタカナが一文字もないツイートは除外した．これら

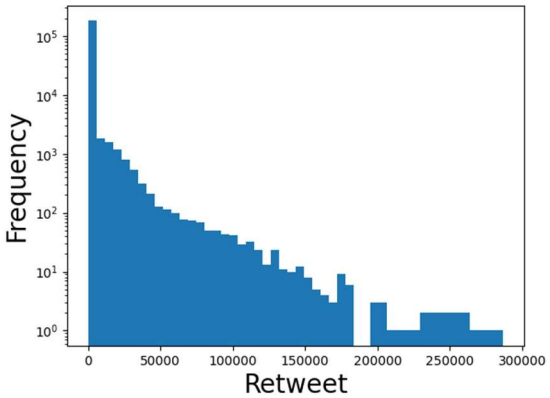

図 1　ファインチューニング用データセットのリツイート数毎の頻度分布

[iii] https://huggingface.co/cl-tohoku/bert-base-japanese
[iv] https://huggingface.co/cl-tohoku/bert-large-japanese
[v] https://developer.twitter.com/en/docs/twitter-api

フィルタ後のリツイート数の分布を図 1 に示す（リツイート数 0 の投稿数も参考に残した）.

要因分析用データセット　上記のまとめアカウントデータのみを用いた．ファインチューニング時に用いたツイートと重複するものは除外した.

表 1 にファインチューニング用と要因分析用のデータセットの統計を示す.

表 1　各データセットの統計情報

	finetuning 用	要因分析用
総ツイート数	121018	4867
最大リツイート数	287012	638670
最小リツイート数	1	15454
平均リツイート数	1710.673	42285.735
最大文字数	276	137
最小文字数	1	6
平均文字数	72.203	103.464

4 実験設定と評価

評価に用いた指標は、平均二乗誤差(MSE)、二乗平均平方根誤差(RMSE)、平均絶対誤差(MAE)、平均絶対パーセント誤差(MAPE)、決定係数(R^2)、相関係数(r)の 6 つである. 評価指標の詳細については付録 A2 に記載する.

4.1 訓練時のデータ分割と評価結果

前章で述べたファインチューニング用データセットを訓練・検証・評価に分割した．ランダムサンプリングしたツイートは 38:1:1、まとめアカウントデータは 8:1:1 に分割した．リツイート数が多いものが訓練・検証・テストそれぞれに適度に含まれるように、それぞれ分割してから結合した．結果、ファインチューニング用のツイート数は、訓練・検証・評価の件数がそれぞれ 113671、3672、3675 となった.

ファインチューニングの結果、ほぼすべての指標において **JTweetRoBERTa** が最も良いスコアを得た（表 2）．表 2 の **follower** カラムはフォロワー数のみで予測した結果を、**JTweetRoBERTa+follower** は **JTweetRoBERTa** の最終全結合層にフォロワー数も入力した場合の予測結果であるが、フォロワー数を加えた結果、かえって性能が低下している．次節でその詳細を分析する.

4.2 リツイート数による性能比較

リツイート数の過多によって予測性能が異なる可能性がある．その調査のため、前節のフォロワー数を使わないモデル間比較で最も良い性能を得た、提案手法である **JTweetRoBERTa** と、**JTweetRoBERTa+follower**、**follower** の三つについて、テストデータをリツイート数 1024(= 2^{10})未満の計 2845 ツイート、2^{10} 以上の計 830 ツイートの二つに分割し、それぞれに対して評価値を算出した．表 3 に結果を示す.

表 3　リツイート分布別の予測性能比較
JTR: JTweet-RoBERTa, f: follower

範囲	$< 2^{10}$			$\geqq 2^{10}$		
モデル	JTR	JTR + f	f	JTR	JTR + f	f
MSE	4.706	5.167	2.622	85.247	90.081	143.489
RMSE	2.169	2.273	1.619	9.233	9.491	11.979
MAE	1.018	0.980	0.842	7.804	8.167	11.896
MAPE	**0.529**	**0.472**	**0.312**	**0.602**	**0.629**	**0.909**
R^2	-1.138	-1.347	-0.191	-42.280	-44.734	-71.849
r	0.105	0.140	0.048	0.138	0.136	0.045

二つのデータセット間でリツイート数のスケールが異なるため、絶対値ではなく比率で評価する MAPE が妥当な指標である．MAPE はリツイート数 2^{10} 未満では **follower** が、リツイート数 2^{10} 以上では **JTweetRoBERTa** が最も良い性能であった.

4.3 文節 Ablation Study による要因分析

リツイート数予測に影響の大きい単語や文節を分析するため、オリジナルの入力ツイート文からいずれかひとつの文節を除いて予測を行い比較した．ツ

表 2　対数正規化リツイート数のモデル別予測の評価値

	Baseline	BERT-large	RoBERTa	JTweet-RoBERTa	JTweet-RoBERTa + follower	follower
MSE	26.211	25.729	26.811	**22.280**	24.907	34.586
RMSE	5.120	5.072	5.178	**4.720**	4.990	5.881
MAE	2.817	2.791	2.986	**2.502**	2.645	3.344
MAPE	0.578	0.616	0.573	**0.570**	0.510	0.449
R^2	-0.080	-0.060	-0.105	**0.817**	-0.027	-0.4256
r	0.403	0.413	0.358	**0.518**	0.446	-0.100

イート本文から削除したときに最も予測値が下がる文節が最も影響を及ぼしていると考え、要因分析用のデータセットすべてについてこの分析を行った.

予測にはファインチューニングした JTweetRoBERTa を使うが、ファインチューニングで用いたものとは別に入力用ツイートデータを用意した. 文節区切りには日本語構文・格・照応解析システム KNP[vi]を用いた. ハッシュタグが KNP の入力に含まれていると正常に動作しないため、一旦ツイート本文からハッシュタグを抽出し KNP に入力した後でツイート本文の元の場所に戻し予測を行った.

表 4 に、ツイート本文をすべて入力した際のリツイート数予測値と、いずれかの文節を抜いた場合のうち最小の予測値とについて、予測値の差上位 10 件をまとめた.

表 4　文節 Ablation Study の分析結果

順位	正解	予測値	最小	差	文節
1	44491	49194	262	48932	お願いなのですが、
2	51678	49686	939	48747	言わせて。
3	24669	48291	1	48290	忍者だった
4	33639	47634	257	47382	一度
5	30726	46553	2	46551	4 年後とか
6	30020	46077	2	46075	聞いてみたら
7	35664	45951	20	45931	【超重要事項】筋肉注射の
8	45932	44418	1544	42874	【拡散希望】脳梗塞って
9	36286	43432	703	42729	聞いてほしいんだけど。
10	37985	42322	163	42159	飼われた

5 考察

表 2 に示した、ファインチューニングしたモデルのうちフォロワー数を使用しないモデル間の性能比較では、提案手法である **JTweetRoBERTa** が最も高い性能であった. ツイート文には特有の語彙、文体、構造があるため大規模ツイートテキストによる事前学習が有効であったと考えられる.

表 3 の **JTweetRoBERTa** の結果を比較すると、リツイート数が多い方が性能が良く、内容以外の要因が相対的に減るであろうこと、リツイート数の多いツイートには予想を決定づける要因が多く含まれたためにより予測を安定させた可能性がある.

リツイート数 1024(= 2^{10})以上のツイートに対する予測で、フォロワー数が貢献しなかったことは意外で重要な発見である. フォロワー数が少ないアカウントからの発信であっても、ツイート内容が「バズる」ものであれば他のアカウントの紹介等を介して最終的にリツイート数が増加するのではないか.

リツイート数 1024(= 2^{10})未満のツイートではフォロワー数のほうが予測に寄与した. フォロワー数が多いほど、内容にかかわらずリツイートするユーザー群が多いのかもしれない.

表 4 の分析結果で挙がった文節を見ると、「お願いなのですが、」「言わせて、」といった相手に働きかけるような文節が上位にみられる. また、トピックをあらわす文節が「忍者だった」「飼われた」「【超重要事項】「筋肉注射の」「【拡散希望】脳梗塞って」のようにアピーリングで意外性のある内容であることも重要であるという結果になった. このことから提案手法では、ツイート本文中の要点に関する文節を重要視してリツイート数予測を行っているのではないかと考えられる.

6 おわりに

本研究では、ツイート本文の言語特徴からリツイート数を予測できるかを試みた. モデル比較の結果、ツイートデータで事前学習させた提案手法の **JTweetRoBERTa** が、リツイート数 1024(= 2^{10})以上の予測において MAPE が 0.602 と最も良い性能を達成した. リツイート数が十分に多い場合フォロワー数の貢献は低くテキストのみで予測が可能であるうえ、Ablation Study の結果からは相手への働きかけ表現やツイート文のトピックを表すアピーリングな文節を重要視しているのではないかという示唆を得た.

今後はファインチューニングに用いたリツイート数の分布がより均一になるよう収集方法を工夫してデータ数を増やし、モデルのさらなる一般化を目指したい. また、ユーザーのツイート履歴やツイートの前後関係、社会背景とツイートタイミングといった時系列の情報など、ツイート本文以外の特徴量も含めたモデルを構築することで、よりユーザーの特性とツイート拡散過程を反映した予測を試みたい.

[vi] version 4.20, https://nlp.ist.i.kyoto-u.ac.jp/?KNP

謝辞

本研究は JSPS 科研費 JP22H00804, JP21K18115, JP20K20509, JST AIP 加速課題 JPMJCR22U4, およびセコム科学技術財団特定領域研究助成の支援をうけたものです.

参考文献

[1] Saurabh Sharma, and Vishal Gupta. Role of twitter user profile features in retweet prediction for big data streams. Multimedia Tools and Applications, Volume 81, pp. 27309–27338. 2022.

[2] Gang Liu, Chuan Shi, Qing Chen, Bin Wu, and Jiayin Qi. A Two-Phase Model for Retweet Number Prediction. Web-Age Information Management. WAIM 2014. Lecture Notes in Computer Science, vol 8485, pp 781–792. 2014.

[3] Qi Zhang, Yeyun Gong, Jindou Wu, Haoran Huang and Xuanjing Huang. Retweet Prediction with Attention-based Deep Neural Network. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 75–84. 2016.

[4] Chunjia Wang, Yongquan Fan, Yajin Du, and Zefen Sun. Predict Individual Retweet Behavior Based on Multi-feature. IOP Conference Series: Materials Science and Engineering, Volume 790, Number 1, pp. 012046. 2020.

[5] Ryo Takasu, Hironobu Nakamura, Taishiro Kishimoto, and Yoshinobu Kano. Mental Health Classification using Large Scale Tweet Dataset. Proceedings of the 36th Annual Conference of the Japanese Society for Artificial Intelligence. 2022.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.0485v2, 2019.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692. 2019.

A 付録

A1 ファインチューン時のパラメータ設定

今回学習に使用した最適化アルゴリズムは Adam である．フォロワー数のみで学習した場合は、最終回帰層のみ学習させた．

lr = 5e-5(各言語モデル)、**1e-5**(最終回帰層)

max_epochs = 100 (early_stopping 設定により早期終了する可能性もある．)

A2 評価指標

本研究で用いた評価指標は、平均二乗誤差(MSE: Mean Squared Error)、二乗平均平方根誤差(RMSE: Root Mean Squared Error)、平均絶対誤差(MAE: Mean Absolute Error)、平均絶対パーセント誤差(MAPE: Mean Absolute Percentage Error)、決定係数(R^2)、ピアソンの相関係数(r)の 6 つである．以下にそれぞれを求める式を示す．

(y: 正解値、$\hat{y}$: 予測値、n: データ数)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{MSE}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$