マイクロブログの再発するトレンドを予測する

赤崎 智　　山下 達雄
LINE ヤフー株式会社
sakasaki@lycorp.co.jp　　tayamash@lycorp.co.jp

概要

我々はマイクロブログにおけるトレンドを日々認識し，最新情報へのキャッチアップを行なっている．これらのトレンドの中には，新しく発生するもののほか，既に発生したものが再度盛り上がる例がある．そこで本稿では，マイクロブログにおけるトレンドの再発を予測するというタスクに取り組む．我々は，マイクロブログである X の過去のトレンドで 2 回以上発生したものを正例，それ以外のものを負例としてデータセットを構築し，どのようなトレンドが再発するかを分析する．予測モデルは，トレンドに含まれるポスト及びユーザプロフィールなどのテキスト情報と，トレンドの勢いなどの時系列情報を用いて再発を予測する．実験では，提案手法がトレンドの勢いのみを特徴量として用いた手法よりも高精度で予測ができることを示す．

1　はじめに

マイクロブログなどのソーシャルネットワークサービスが実世界のありとあらゆる情報を素早く拡散し，広範な観点からの意見を伝達する手段として一般的に使用されるようになって久しい．

このようなサービスでは，特定のトピックが短期間で大規模な注目を集める現象，いわゆる「トレンド」が発生する．これらのトレンドは，個人の趣味嗜好や企業のマーケティング戦略を形成するなど，多岐にわたって社会に影響を及ぼす力を持っており，地域や世界的なニュース，特定の作品，人々や団体に関する情報，ミームに至るまで多種多様なトピックを含んでいる．

このようなトレンドを早期に予測することは，ユーザへの話題提供のみならず，企業活動における戦略策定や広告のタイミング調整に有効であり，政策立案者にとっても，公衆の関心の移り変わりを理解し，それに応じた世論把握や国民への情報提供を行うための重要な手段となりうる．

トレンド予測に関しては多くの取り組み [1, 2, 3, 4, 5, 6] があり，これらは基本的には新規に発生するトレンドのみを扱っているが，実際には過去に一度トレンドになったトピックが再び浮上するケースも頻繁に観測される．このようなトレンドの再発を予測することは，将来大きくヒットするトピックの推測や，ユーザの行動パターンの予測など，多くの課題に役立つ可能性がある．

そこで本稿では，マイクロブログにおけるトレンドの再発を予測するという新しいタスクを提案する．我々は本タスクに取り組むにあたり，マイクロブログとして X（旧 Twitter）を対象とし，X の過去のトレンドを収集しトレンドの再発を予測するためのデータセットを構築した上で，どのようなトピックがトレンドとして再発しやすいかを分析する．更に，構築したデータセットを用いて，実際に与えられたトレンドが再発するかどうかを，ポストやユーザ情報，トレンドの勢いなどといった特徴量を用いた機械学習によって予測し，結果を分析する．

2　関連研究

X などのソーシャルメディアのトレンド予測に関する研究は数多く存在する [1, 2, 3, 4, 5, 6]．これらの研究では，トレンドの初発に関する予測分析が試みられ，ユーザ行動，時系列頻度，テキスト，ネットワーク構造などの様々な特徴量を用いて予測を行っている．しかしながら，既に発生したトレンドが再度注目を集めるかや再発するかどうかについては考慮されていない．

これに対して，Cheng ら [7] はソーシャルメディア上で共有されるコンテンツの人気が突発的に再浮上する現象に着目し，それらが初期の拡散状況からある程度予測可能であることを明らかにしている．しかしながら，彼らの研究はそのままリポストされやすい画像や動画コンテンツのみを対象としているため，マイクロブログの多様なトレンドがどのように再発するかは明らかとなっていない．

一方，ソーシャルメディアのトレンドに影響を与える要素として，情報拡散のダイナミクスについての研究が存在する [8, 9]．これらの研究では，情報がネットワーク間でどのように伝播するか，また，特定の情報がどのようにトレンドを生み出すかを分析している．しかし，これらの研究でもトレンドの再発の予測については着目されていない．

本稿では，トレンドの再発という現象に着目し，データセットの構築および分析を行い，トレンドの初期のポストやメタデータを用いて再発を予測するというタスクに取り組む．

3 データセット

本節では本稿で扱う X のトレンドと，再発するトレンドを予測するタスク及びデータセットを構築する手順について説明する．

3.1 トレンド

X では，関心の高い話題が「トレンド」としてリスト化される．これらは，短時間で大量のポストが投稿されることを示し，社会的な注目の焦点や人々の関心を反映している．トレンドは地理的位置，時間，特定のイベントなどによって，時々刻々と変化し，例によっては一度減退したものが，再び社会的関心が高まるときに再度活性化することもある．X のトレンドは，投稿頻度などを考慮した独自の非公開アルゴリズムによって決定される．[1)]

3.2 タスク設定

本稿では，Twitter のトレンドの初期段階，すなわちトレンドが発生した日付の情報を用いてそのトレンドが再発するかどうかを予測するというタスクを設定する．これは，速やかな対応が求められる世論把握やマーケティングなどの応用上の観点から重要であり，また再発するトレンドの中には，再発するまでの間隔が数日以内の短いものも存在するためである．関連研究においてもトレンド初期の情報がトレンドの動向の手がかりとなると報告されており [5, 7]，初期段階での予測は高精度なトレンド再発予測を実現する上で有効であることを示している．

また，トレンドの発生のタイミングや規模については，その予測が困難であることが関連研究で述べられている [7, 10]．そのため，本稿では現実的な問題設定として，与えられたトレンドが再発するか否かの二値分類問題として予測タスクを設定する．

3.3 データセットの構築

まず，X での各日付の日本語のトレンド上位 50 件を，アーカイブサイト[2)]から収集した．我々は，このうちの 2020 年および 2021 年のトレンドを用いてデータセットを構築する．まず，これらのトレンドから 2019 年以前に出現したものを除去し，すでに再発した可能性のあるトレンドがデータセットに含まれないようにした．そして，各トレンドについて，発生した日付の 2 日後以降[3)]から 2023 年 12 月までに再発した例を正例とし，その他のトレンド，すなわち再発していないものを負例として自動的にラベルを割り振った。

次に各トレンドに付随するポストとして，そのトレンドが発生した日を時間で区切り，トレンド名が最も多くポストされた時刻のものをランダムに最大 1,000 件収集した．表 1 に収集した結果を示す．

3.4 データセットの分析

表 2 は，データセットの再発したトレンドについて，それが再発するまでの日数ごとに分けたものを示す．表より，トレンドが再発するまでの日数は，数日程度の短いものから数年以上の長いものまで幅が広いことがわかる．数日程度の短いものとしては，「クルーズ船」（コロナウィルスの患者が発生したダイヤモンドプリンセス号）のような短期間の動向が注目されやすいトピックであったり，「#呪術廻戦」のような毎週放映され X 上で実況される人気アニメーションのトピックなどもあった．それより長い中期間程度のものとして，「かぐや様 3 期」や「#えんとつ町のプペル」などがあった．これらのような作品のトピックはその制作が発表され公開に至るまでの過程，日数としては数十日以上から 1 年未満の期間で度々注目を浴び再度トレンドとなることが多い．長期間，すなわち数年以上の長いものについては，年単位で発生・開催されるイベントなどのほか，昔のミームやハッシュタグなどがなんらかのきっかけで再び関心を浴びる例があった．後者についてはポストの内容にも脈絡がなく，正確な再発の予測は難しいと考えられる．

1） https://help.twitter.com/en/using-x/x-trending-faqs

2） https://jp.trend-calendar.com/trend/2020-01-01.html

3） 発生した 1 日後だとトレンドが再発ではなく継続している例があるため

年	再発		単発	
	トレンド数	ポスト数	トレンド数	ポスト数
2020	1,557	914,171	7,759	4,002,916
2021	1,386	783,960	7,756	3,590,844

表 1 収集したトレンドとポスト

再発までの日数 (d)	トレンド数	トレンド例	ポスト例
$2 \leq d \leq 4$	58	クルーズ船, オミクロン株	横浜のクルーズ船の人達どうなんの？ 一生隔離？
$4 < d \leq 8$	133	#呪術廻戦, #バチェラー 4	呪術アニメ始まった〜 #呪術廻戦
$8 < d \leq 16$	146	チェンソーマン, 一気読み	チェンソーマン読んだ。マキマさんヤバすぎる。 URL
$16 < d \leq 32$	210	一粒万倍日, #非同期テック部	24 日は一つのことが万倍になる縁起が良い一粒万倍日
$32 < d \leq 64$	296	ハケンの品格, 世の中 4 連休	世の中 4 連休らしいですが、私は月曜まで普通に仕事〜
$64 < d \leq 128$	388	かぐや様 3 期, #えんとつ町のプペル	かぐや様 3 期やるのめちゃ嬉しい〜
$128 < d \leq 256$	439	アルカンタラ, クソデカ林製薬	朝からアルカンタラ選手獲得！？ビックニュースやん！
$256 < d \leq 512$	751	クリぼっち, マリカ杯	今年もマリカ杯あるのやっちゃ！
$512 < d \leq 1024$	462	所得制限, マスク着用不要	現金給付より消費税減額の方が良くないですかね？
$1024 < d \leq 2048$	60	ペルセウス座流星群, #日本グミ協会	これがペルセウス座流星群なの？ URL

表 2 収集したトレンドを再発するまでの日数ごとに分けたものとポスト例（実例より再構成したもの）

4 実験

本節では構築したデータセットを用いて教師あり分類器を構築し，再発トレンドの予測を試みる．

4.1 特徴量

再発するトレンドをモデリングするため，以下の特徴量を導入する．

時系列頻度: トレンドに付随するポストの時系列頻度．具体的には，トレンドとなった日付において，頻度の最大勾配，平均勾配，最大頻度，最大頻度の時刻前と後における平均勾配を計算し用いる．

ポスト: トレンドに付随するポストの内容．ポストの数は膨大で全てを用いるのは困難であり，トレンドのトピックも広範なため，単純に Bag-of-words などを用いると特徴量ベクトルが疎になり次元数も巨大になってしまう．これに対し，近年自然言語処理の様々なタスクで高い性能を挙げている文埋め込みを利用することで問題の緩和を図る．具体的には，トレンドの各ポストの文埋め込みを獲得し，それらの平均を取り特徴量として用いる．

ユーザプロフィール: トレンドに付随するポストを投稿したユーザのプロフィール．トレンドごとに，**ポスト**で用いたポストのユーザのプロフィールの文埋め込みを獲得し，これらの文埋め込みの平均を取ったものを特徴量として用いる．

4.2 比較手法

以下の手法を用いてトレンドが再発するか否かの分類を行い，性能を比較する．

AllPositive: 全てのトレンドについて「再発」ラベルを出力する手法．

TemporaryOnly: 時系列頻度のみを用いて機械学習モデルを学習し，分類を行う手法．

Proposed: 全ての特徴量を用いて機械学習モデルを学習し，分類を行う手法．

4.3 設定

特徴量の文埋め込みを獲得するため，2018 年 4 月から 2019 年 4 月の期間にランダムに X からサンプルしたポスト 5,000 万件を用いて最大文長 64，バッチ数 16 に設定した BERT [11]（bert-base）を 40 エポック事前学習した．更に，それを用いて対照学習により文埋め込みを獲得する手法である教師なし SimCSE [12] を標準のパラメータで 1 エポック学習した．ポストのトークナイズには sentencepiece[4] を用い，最適化には Adam を用いた．

機械学習モデルとして，決定木とアンサンブル学習を組み合わせた勾配ブースティング木 [13] を用いた．実装としては Python と lightGBM[5] を用い，optuna[6] の LightGBM Tuner で最適化を行った．

学習及び評価を行うため，2020 年のデータを学習データ，2021 年のデータを評価データとして用いた．学習データはそのまま使うと負例の数が正例より極端に多くなってしまうため (表 1)，アンダーサンプリングを用いて負例を正例の数と同数に揃えるようにした．開発データは学習データの 1 割を用い，開発データで最もロスが低いモデルを評価デー

4) https://github.com/google/sentencepiece
5) https://lightgbm.readthedocs.io/en/stable/
6) https://optuna.org/

	Accuracy	Precision	Recall	F_1-score
AllPositive	15.16	7.58	50.00	13.16
Temp.Only	58.09	53.96	57.59	49.70
Proposed	**64.32**	**56.45**	**61.85**	**54.40**

表 3　再発トレンド分類の手法ごとの結果比較

	Precision	Recall	F_1-score	Support
再発しない	89.77	65.39	75.67	7756
再発	23.14	58.30	33.13	1386

表 4　**Proposed** の再発トレンド分類の結果

タに適用した．10 回平均したものを最終的な評価値として用いた．

4.4　結果

表 3 に，評価データである 2021 年のトレンドについて二値分類を行った結果を示す．表より，時系列頻度のみを用いた場合より，ポストやユーザプロフィールなどのテキスト情報も用いた方が精度が向上することがわかる．トレンドの勢いや人気度合いは過去のトレンド予測の研究でも精度に寄与することが報告されているが，テキスト情報との組み合わせがより有効であることがわかる．

表 4 は提案手法の各ラベルの精度を示している．表より，再発しないトレンドは比較的精度良く予測できているものの，再発するトレンドについては精度が悪いことがわかる．特に Precision が低く，現状の手法や特徴量では再発するトレンドについてのモデリングが十分でないことが伺える．

表 5 は，評価データについて，再発までの日数ごとにデータを分けた際の Recall を示している．ほとんどの日数で概ね 6 割程度の数値となっているが，短期間の $2 \leq d \leq 4$ 及び中期間の $32 < d \leq 64$ と $64 < d \leq 128$ はそれを下回っており，これらの例は再発を判定するのがより難しいことがわかる．特に最も数値が低い $2 \leq d \leq 4$ のような例は，重大なニュースなどの数日以内に再び注目を浴びるトピックを適切に認識する必要がある．$4 < d \leq 8$ の数値は最も高く，この期間には 3.4 節で述べたような毎週放送される作品などが多く含まれており，そのような例の判定は比較的容易であると考えられる．

最後に，評価データの正例に対する **Proposed** の予測例をいくつか挙げる．**Proposed** は，「ハイパーインフレーション」，「水星の魔女」などの作品系のトレンドの再発を正しく予想できていた．前述した通り，このような例はポストに「更新」，「連載」，

再発までの日数 (d)	トレンド数	Recall
$2 \leq d \leq 4$	33	48.48
$4 < d \leq 8$	60	68.67
$8 < d \leq 16$	62	66.45
$16 < d \leq 32$	102	62.16
$32 < d \leq 64$	130	54.62
$64 < d \leq 128$	202	54.36
$128 < d \leq 256$	217	60.74
$256 < d \leq 512$	397	59.60
$512 < d \leq 1024$	183	59.23

表 5　再発までの日数ごとに分け Recall を確認した結果

「放映」，「来週」など，周期的な盛り上がりを予感させる単語を多く含んでおり，これらを捉えることで再発を予測することができる．一方で，「ルームマッチの対戦相手」，「#ウチカフェしよう」などの特定のイベントに関するトレンドや，「#一番為になった PC 知識」のような突発的なトレンドは，ポストなどを人目で確認しても将来再度盛り上がるかが不透明であり，再発の予測が難しいことがわかった．

5　まとめ

本稿では，マイクロブログにおけるトレンドが再発するという現象に着目し，実際の X のトレンドを収集した上でそれらが再発するか否かのラベルを自動的に割り振りデータセットを構築した．データセットを分析した結果，トレンドの再発は幅広い期間で起き，期間によっては特有のポストを伴うことを確認した．我々は構築したデータセットを用いてトレンドの再発を予測するタスクを設定し，特徴量を設計した上で実験を行った．実験の結果，精度には改善余地があるが，特定の期間に再発するトレンドに関しては予測がしやすいなどの知見を得た．

今後の予定としてはまず予測精度の改善である．今回はトレンドが発生した日付のみから特徴量を抽出しているが，その日付以降の時系列も用いることでよりトレンドの振る舞いを捉えることができると考えられる．また，データ収集についても予測が困難なノイジーなトレンドの除去や，予測に有用なポストの選別などに取り組む予定である．実用面では，弊社のマイクロブログ検索・トレンド抽出サービスである「Yahoo!リアルタイム検索[7]」への適用を進め，研究成果の社会還元を目指す予定である．

7)　https://search.yahoo.co.jp/realtime

参考文献

[1] Gabor Szabo and Bernardo A Huberman. Predicting the popularity of online content. **Communications of the ACM**, Vol. 53, No. 8, pp. 80–88, 2010.

[2] Jiang Yang and Scott Counts. Predicting the speed, scale, and range of information diffusion in twitter. In **Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)**, pp. 355–358, 2010.

[3] Roja Bandari, Sitaram Asur, and Bernardo Huberman. The pulse of news in social media: Forecasting popularity. In **Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)**, pp. 26–33, 2012.

[4] Suman Deb Roy, Gilad Lotan, and Wenjun Zeng. The attention automaton: Sensing collective user interests in social network communities. **IEEE transactions on network science and engineering**, Vol. 2, No. 1, pp. 40–52, 2015.

[5] Benjamin Shulman, Amit Sharma, and Dan Cosley. Predictability of popularity: Gaps between prediction and understanding. In **Proceedings of the international AAAI conference on web and social media (ICWSM)**, pp. 348–357, 2016.

[6] Shogo Matsuno, Sakae Mizuki, and Takeshi Sakaki. Construction of evaluation datasets for trend forecasting studies. In **Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)**, pp. 1041–1051, 2023.

[7] Justin Cheng, Lada A. Adamic, Jon M. Kleinberg, and Jure Leskovec. Do cascades recur? In **Proceedings of the 25th International Conference on World Wide Web (WWW)**, p. 671–681, 2016.

[8] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In **Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)**, pp. 137–146, 2003.

[9] Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A Zighed. Information diffusion in online social networks: A survey. **ACM Sigmod Record**, Vol. 42, No. 2, pp. 17–28, 2013.

[10] Fan Zhou, Xovee Xu, Goce Trajcevski, and Kunpeng Zhang. A survey of information cascade analysis: Models, predictions, and recent advances. **ACM Computing Surveys**, Vol. 54, No. 2, 2021.

[11] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In **Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (NAACL-HLT)**, pp. 4171–4186, 2019.

[12] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, 2021.

[13] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (KDD)**, pp. 785–794, 2016.