

# ソーシャルメディアにおける一時的な出来事の検出

服部 寛之<sup>†</sup>    鈴木 優<sup>†</sup>

<sup>†</sup> 岐阜大学工学部電気電子・情報工学科 〒501-1193 岐阜市柳戸 1-1

E-mail: <sup>†</sup>{w3033111@edu.gifu-u.ac.jp, ysuzuki@gifu-u.ac.jp}

**あらまし** 本研究の目標は社会での変化や出来事の発生を、ソーシャルメディアのデータを基に自動的に提示することである。これにより、過去に発生した出来事の振り返りが可能になり、将来に備えることが出来るようになる。我々は社会での変化や出来事に関するデータを、教師あり学習を用いて自動的に集めようとしているが、モデルを構築するためのデータが必ずしもあるとは限らない。そこで、短期間に発生した出来事とは直接関係のないデータを学習させて作ったモデルの分類精度に着目して、一時的に発生した出来事の発生を推定できるのではないかと考えた。提案手法では、ツイートの投稿日時を区間に分割した後、それぞれの区間で分類精度を算出する。ある区間の分類精度は、その区間に投稿されたツイートを、それ以外の区間に投稿されたツイートを学習した機械学習のモデルに入力して求める。実験では、提案手法を用いて、区間の分類精度と一時的な出来事の発生件数には相関があるか調べた。提案手法により、既存のラベル付きのデータの活用方法が広がることが期待されるほか、ラウドソーシングの作業に適用させて、少ない労力で、短期間に発生した変化や出来事を見つけることが可能になる。

**キーワード** イベント検出, Twitter, 機械学習

## 1 はじめに

本研究では社会での変化や出来事の発生を、ソーシャルメディアのデータを基に自動的に提示することを目的とする。これにより、過去に発生した出来事の振り返りが可能になり、将来に備えることが出来るようになる。しかし、大量の情報から社会の変化や出来事を網羅的に発見することは難しい。なぜならクラウドソーシングを使ってソーシャルメディアのデータから社会での変化や出来事を見つけることは作業内容が簡単ではない上に、金銭的なコストがかかるからである。そこで、本研究では機械学習を用いて、短期間に発生した出来事の発生を検出する方法を提案する。

ここで本稿では、ある出来事が発生し、終了するまでの期間が短い出来事かつ、繰り返し発生していない出来事を一時的な出来事と呼ぶ。一時的な出来事と呼ばれる例の一つを挙げると、2020年2月から3月の変わり目にかけてトイレットペーパーの買い占めが起き、品薄になっているというデマ情報が発生した出来事が挙げられる。この出来事は約一週間という短い期間で収まり、同様のデマ情報が繰り返し発生することはなかった。そのため、一時的な出来事であると判断できる。また、一時的な出来事ではないという例を挙げると、新型コロナウイルスの一日の新規陽性者数が2020年4月8日に500人を超えた出来事である。この出来事は、一日という短い期間で発生したが、新型コロナウイルスの陽性者数は毎日公表されている。そのため、出来事が繰り返し発生しているとみなす。従って、一時的な出来事ではないと判断する。

一時的な出来事の発生をソーシャルメディアの大量のデータを基に、自動的に検出を行うには、まず、教師あり学習を利用して、一時的な出来事に関するデータを集めることを行う。し

かし、一時的な出来事であるかどうか分類したデータセットを著者らが探したところ、見つけることができなかった。そのため、この方法をとることができない。また、新たにデータセットを作るとしても、費用や作業が要する。そこで、我々は一時的な出来事とは直接関係のないデータを学習させて作った分類器の分類精度に着目して、一時的な出来事の発生を推測できないかと考えた。目的とは直接関係のないデータセットを学習させて構築した分類器の分類精度と、一時的な出来事の件数に相関があれば、その分類器を利用することで、一時的な出来事の発生を検出することが出来るのではないかと考える。

本研究ではソーシャルメディアとして利用者の多い、Twitterのツイートを利用した。提案手法では、まず、ツイートの投稿日時をもとに区間を設定する。具体的には、ツイートの投稿日時をいくつかの区間に分ける。その後、それぞれの区間で分類精度を算出する。ある区間の分類精度は、その区間に投稿されたツイートを、それ以外の区間に投稿されたツイートを学習した分類器に入力して求める。そして、設定した区間のうち、分類精度が低い区間で一時的な出来事が発生しといるとみなす。我々が分類精度が低い区間に着目する理由は、設けた区間のうち、1つの区間にだけ、ある出来事に関するツイートが投稿されている場合、その出来事に関するツイートの学習が不十分のために、分類は不正確となりやすいと考えるからだ。

実験では、提案手法を用いて、区間の分類精度と一時的な出来事の発生件数には相関があるか調べた。分類精度を区間ごとに算出するためのラベル付きツイートとして、コロナウイルスに関係しているツイートを事実と感想に分けたツイートを利用した。また区間の設定は、分割する区間の長さを長くする場合と、短くする場合の二通りを用意した。一時的な出来事の発生

件数は、著者らがNHKの新型コロナウイルス関連の記事<sup>1</sup>から作成した。機械学習のモデルとしてBERTを用いた。また、分類精度の評価指標としてAccuracyを利用した。結果は、分割する区間の長さを長くする場合では、分類精度と一時的な出来事の発生件数には相関がみられなかった。しかし、分割する区間の長さを短くする場合では、負の相関がみられた。具体的には一時的な出来事が一件発生するごとに、分類精度は0.021下がった。

本研究の貢献は、人手によって、一時的な出来事とは無関係なラベル付けがされたソーシャルメディアの大量のデータから、自動的に短期間に発生した出来事の検出方法を提案したことである。これにより、既存のラベル付きデータの活用方法が広がることが期待されるほか、クラウドソーシングの作業に適用させて、少ない労力で、短期間に発生した変化や出来事を見つけることが可能になる。

## 2 関連研究

ツイッターに投稿されるツイートを観測し、イベント発生の検出やイベントの詳細な情報を得るという試みが行われている。ツイッター上でのイベントの発生の検出は主に、バーストと呼ばれるツイートの急激な増加やある単語の急激な増加など、データストリームの異常な状態を検出する方法が用いられている。

羽山ら[1]はテレビ番組のダイジェスト自動生成のためのイベントの同定のために、特定のユーザーのグループが投稿したツイートに対してKleinberg[2]の方法を用いている。また、坂本ら[3]はTwitterストリームの自動要約のためのイベント検出を行うために、特定のハッシュタグを含むツイートに対して蝦名ら[4]の提案したリアルタイムバースト検出手法を用いている。いずれのイベント検出もツイートの増加に注目している。

また、Sakakiら[5]は台風や地震などのイベントの検出のために、イベントに関係するキーワードを含むツイートを収集し、集めたツイート数の時間的な分布からイベントの発生確率を算出してイベントの検出を行った。しかし、Sakakiらは羽山らや坂本らの方法とは異なり、イベント検出を行う前に、キーワードから集めたツイートの中から関係のないツイートを取り除くということも行っている。具体的には、収集したツイートが実際に発生したイベントについて言及しているかどうかの判定をSVM(Support Vector Machine)を用いて行っている。

また、イベントの発生を検出した後、観測したツイートに含まれる情報からイベントの詳細な内容を得るという試みも行われている。羽山ら[1]、坂本ら[3]はツイートのバースト検出を行った後、バーストしているツイートに使われている単語に着目している。特に、羽山らはバーストが検出された時間帯において、さらに単語のバースト検出を行っている。そして、バーストしている単語から、イベントの内容を推測している。

水沼ら[6]は、ツイッター上でのバーストの形質がイベントの性質を表していると考え、バースト時の継続時間やリツイ

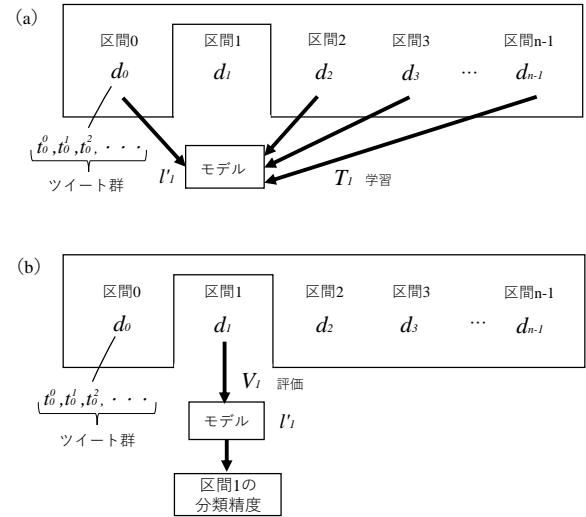


図1 分類器の構築方法と分類精度の算出方法

トの比率、ツイート本文の文字数などを特徴量として類型化し、類型ごとにイベントの性質について推測を行った。

また、Sakakiら[5]はイベントの検出後、イベントが発生した場所の推定を行うために、プロフィールにユーザーの場所が登録しているユーザーの投稿ツイートや、位置情報が付けられたツイートを利用してイベントの発生場所を特定している。

本研究が関連研究と異なる点は、本研究はイベント検出において、教師あり学習のツイートの分類精度に着目した点である。また、一時的な出来事の検出を行う際に、教師あり学習の学習用データには、一時的な出来事に関するツイートとは直接的に関係のないラベルが付与されたツイートを用いたことである。

## 3 提案手法

本研究では一時的な出来事の検出の試みとして、機械学習と一時的な出来事とは無関係なラベルが付いたツイートを利用する。一時的な出来事の検出は、分類精度が低い区間に着目して行う。一時的な出来事の検出のための、分類器の構築及び分類精度の算出は区間ごとに行う。区間の設定は、ツイートの投稿日時を分割して行う。本研究では、分類器の構築と分類精度の算出方法が一般的な機械学習の分類器構築と分類精度の算出方法と異なる。提案手法の分類器の構築の方法と分類精度の算出方法について図1に示す。分類器の構築と分類精度の算出方法の詳しい説明は、3.3節と3.4節で行う。この図では、区間1の分類器の構築と分類精度の算出を行っている様子が示されている。図中の(a)では、まず、分類器構築のために、区間1以外に属しているツイートを機械学習のモデルに学習させている。その後、図中の(b)で、構築した分類器に区間1に属しているツイートを評価用データとして投入し、区間1の分類精度を算出する。

提案する分類器の構築と分類精度の算出の方法では、ある区間で投稿されたツイートの内容が、別の区間でも投稿されていると、その内容のツイートの分類は学習量が十分であるため、

<sup>1</sup> : <https://www3.nhk.or.jp/news/special/coronavirus/chronology/>

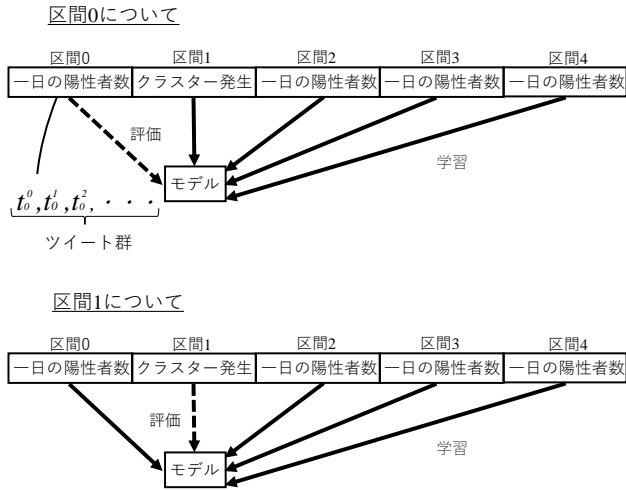


図 2 区間の分類精度算出の例

分類は正しく行われると考える。一方、ある区間で投稿されたツイートの内容が、それ以外の区間で投稿されていないと、その内容のツイートの分類は、学習量が少ないため、不正確となりやすいと考える。例えば、図 2 の場合では、区間 0、及び区間 2 から区間 4 に属しているツイートは、その大半が新型コロナウイルスの一日の陽性者数に関するツイートとする。一方、区間 1 に属しているツイートの大半は、ある場所で発生したクラスター（集団感染）に関するツイートであるとする。この場合、区間 0 の分類精度を算出するためのモデルには、新型コロナウイルスの一日の陽性者数に関するツイートが比較的多く学習されるため、区間 0 の分類精度は比較的高いと考えられる。一方、区間 1 の分類精度を算出するためのモデルには、クラスター発生に関するツイートは殆ど学習されないため、区間 1 の分類精度は低いと考えられる。従って、他の区間で発生していない出来事が起きている区間では、分類精度が低いと考えられる。言い換えると、一時的に他の区間で発生していない出来事が起きている区間では、分類精度が低くなると考えられる。

一時的な出来事が発生した区間を検出する手法の流れを次に示す。

- (1) ツイートの投稿日時を区間に分割する。(3.1 節)
- (2) ラベル付けが行われたツイートを区間別集める。(3.2 節)
- (3) 区間ごとに分類器の構築及び分類精度の算出を行う。(3.3 節, 3.4 節)

### 3.1 区間の設定

我々の提案手法では、分類精度の算出は設定したそれぞれの区間で行う。区間の設定方法によって区間の数や区間の長さが変わる。区間の設定方法は様々考えられるが、我々は区間の設定方法として、区間の長さを揃える方法とツイートの数を揃える方法の二つを提案する。

#### 3.1.1 区間の長さを揃える方法

区間の長さを揃える方法は、ツイートの投稿日時をもとに、区間を一定の期間ごとに分けるやり方である。この方法では、例えば、区間を一時間ごと、一日ごと、一週間ごと、一箇月ごとのように設定する。

この方法でのメリットは、すべての区間の長さが等間隔になるため、区間の分け方が感覚的に分かりやすく、区間を簡単に決めることができるという点である。しかし、この方法では、ラベル付きのツイートの収集後、各区間でラベル付きツイートの数が異なるというデメリットがある。

#### 3.1.2 ツイートの数を揃える方法

ツイートの数を揃える方法は、ラベル付きツイートの数を、各区間で等しく揃えるように区間を分けるやり方である。具体的な方法としては、ラベル付きツイートの収集後、それらを投稿日時順に並べ、あらかじめ定めたツイートの個数ごとに区間を順に分ける。例えば、1 区間あたりのツイート数を 600 と決め、ラベル付きツイートを 12,000 件収集できたとする。この場合、ラベル付きツイート 12,000 件を投稿日時順に並べた後、600 件ずつ順にツイートを分けて行う。なお、この例の場合、区間の総数は 20 となる。

この方法でのメリットは、すべての区間において、ラベル付きツイートの数が等しくなる点である。しかし、この方法では、区間は一定の期間ごとに設定されていないため、区間の長さが各区間で異なる。

区間の設定において、区間を  $n$  個設け、時系列順に区間を  $s_0, s_1, s_2, \dots, s_{n-1}$  と名付けたとき、区間の集合  $S$  は次のように表される。

$$S = \{s_0, s_1, s_2, \dots, s_{n-1}\}$$

### 3.2 データの収集

区間ごとにラベル付けがなされたツイートを集める。収集するツイートの件数は各区間それぞれ可能な限り大量に集めるものとする。ラベル付きのツイートの収集後、各区間でラベルの比率を等しくする。その理由は収集したツイートのラベルの数の大きな偏りがあると、一時的な出来事が発生している区間を正しく検出することができないと考えられるからである。本手法ではラベルの比率が異なる場合は剰余のツイートを取り除く。

ある区間  $s_k$  に属す  $N$  個のツイート  $t_{s_k}^0, t_{s_k}^1, t_{s_k}^2, \dots, t_{s_k}^{N-1}$  の集合  $d_{s_k}$  は次のように表される。

$$d_{s_k} = \{t_{s_k}^0, t_{s_k}^1, t_{s_k}^2, \dots, t_{s_k}^{N-1}\}$$

### 3.3 分類器の構築方法

分類器の構築と分類精度の算出方法が一般的な機械学習の方法と異なる点は、収集したデータの取り扱いが、一般的な機械学習のものとは異なる点である。具体的には、収集したデータを学習用のデータと、評価用のデータに分けるときの分け方が異なる。学習用のデータと、評価用のデータの定め方を数式 (1) と (2) に示す。

なお、我々の提案手法では、分類精度の算出は区間ごとに行

う。また、分類精度を求めるための、分類器の構築は、それぞれの区間で行う。つまり、一時的な出来事が発生している区間の検出のために、 $n$  区間設けた場合、分類器は合計  $n$  個作ることになる。

ある区間  $s_k$  の学習用のツイート集合を  $T_k$ 、評価用のツイート集合を  $V_k$  とすると、 $s_k$  の学習用データ及び評価用データは次のように表される。

$$T_k = \bigcup_{i=0}^{n-1} d_{s_i} \setminus d_{s_k} \quad (1)$$

$$V_k = d_{s_k} \quad (2)$$

$T_k$  は、ツイートの全体集合から、区間  $s_k$  に属しているツイートの集合を除いたものである。 $V_k$  は、区間  $s_k$  に属しているツイートの集合である。

なお、分類器には BERT [7](Bidirectional Encoder Representations from Transformers) を利用する。BERT は事前学習を行ったモデルにファインチューニングを行うことで様々な自然言語処理のタスクで高い精度の文章分類が可能となっている。分類器の構築において BERT を使った理由は、様々な文章を用いたタスクにおいて分類精度が高いためである。

BERT の事前学習済みのモデルは東北大学の乾・鈴木研究室が公開している訓練済み日本語 BERT モデル<sup>2</sup>を用いる。分類器の構築は事前学習を行ったモデルに分類用のレイヤーを追加し、学習用データで、分類用のレイヤーと BERT の最終層に対して、ファインチューニングを行って実現する。

学習を止めるタイミングは 1 エポックごとに評価用データの損失関数の値を算出し、直近 10 エポックの学習において損失関数の最低値が一度も更新されなときとした。区間の分類精度として採用した Accuracy は各エポックの中で最も損失関数の値が低いときのものとした。また、学習用データ  $T_k$  によって構築した分類器を  $l'_k$  とする。

### 3.4 分類精度の算出方法

ある区間  $s_k$  の分類精度の算出は、分類器  $l'_k$  に、評価用データ  $V_k$  を入力して行う。分類精度として評価する指標は構築した分類器に評価用データを入力したときの Accuracy とする。

ある区間  $s_k$  の評価用データの分類精度  $Acc(s_k)$  は次の式で表される。

$$Acc(s_k) = \frac{|\{t_{s_k}^m | l(t_{s_k}^m) = l'_k(t_{s_k}^m), 0 \leq m < N\}|}{|d_{s_k}|}$$

この式の右辺の分母は、ある区間  $s_k$  に属しているツイート数、すなわち、区間  $s_k$  の評価用データの数を表す。また、分子は区間  $s_k$  の評価用データのうち、正解ラベルと分類器による分類が一致した数を表している。 $l(t)$  はツイート  $t$  に付けられた正解ラベルであり、 $l'_k(t)$  はツイート  $t$  を分類器  $l'_k$  に入力したときの、分類器の出力である。

## 4 実 験

本稿では二つの実験を行った。一つ目の実験では、各区間のデータ数のばらつきが分類精度に影響を与えているのかを調べた。また、二つ目の実験では、提案手法の 3.1 節から 3.4 節までを行い、分類精度と一時的な出来事の発生件数の相関を調べた。

### 4.1 実 験 1

この実験では、各区間のデータ数のばらつきが分類精度に影響を与えているのかを調べた。実験を行った理由は、各区間のデータ数のばらつきが、各区間の学習用データ数が不揃いになる原因を作るからである。各区間の学習用データ数が不揃いになると、分類精度は学習用データ数によって左右される可能性があり、一時的な出来事が発生している区間の誤検出につながる恐れがある。従って実験を行い、区間のデータ数のばらつきが分類精度にどのくらい影響を与えているかについて調べた。なお、データの個数が少ない区間ほど、学習用データ数が増えるため、実験の結果として、1 区間のデータ数が少ない区間は、Accuracy が高くなることが事前に予想される。

#### 4.1.1 実験方法

各区間でデータ数を統一することはせずに、提案手法の 3.1 節から 3.4 節までを行った。そして、各区間の Accuracy とデータ数をペアで取得した。その後、取得したペアから Accuracy とデータ数の相関係数を求めた。負の相関がみられる場合は、回帰直線を求めてデータ数が増えたときの Accuracy の減少割合を調べる。

実験に用いたラベル付きツイートは、コロナウイルスに関係しているツイートが、事実であるか感想であるかラベル付けがされたツイートを用いた。これらのツイートは COVID-19 日本語 Twitter データセット<sup>3</sup>から収集を行った。このデータセットは「COVID」もしくは「コロナ」を含むツイートに対し、新型コロナウイルスに関係するツイートか否かの判定を 5 名～10 名の作業者に依頼し、作業者それぞれの判定を多数決でまとめた結果が載せてある。そのデータセットには、そのほかにも新型コロナウイルスに関係するツイートが事実であるか感想であるか判定付けされた結果も載せている。我々はそのデータセットのなかのコロナウイルスに関係しているツイートが、事実であるか感想であるかラベル付けされたツイートを用いた。その理由は、我々は新型コロナウイルスに関係する出来事で、一時的な出来事が発生した区間の検出を試みたことと、区間の総数と 1 区間当たりのデータの数をできる限り大きくするためである。

データの収集のため、はじめに COVID-19 日本語 Twitter データセットからコロナウイルスに関係するツイートで事実であるとラベル付けされたツイートの ID と、コロナウイルスに関係あるツイートで感想であるとラベル付けされたツイートの ID を取得した。その後、取得したツイートの ID からツイート

2 : <https://github.com/cl-tohoku/bert-japanese>

3 : [http://www.db.info.gifu-u.ac.jp/data/Data\\_5f02db873363f976fce930d1](http://www.db.info.gifu-u.ac.jp/data/Data_5f02db873363f976fce930d1)

表 1 各区間のデータ数					
区間	事実	感想	区間	事実	感想
1月上旬	25	25	4月中旬	197	197
1月中旬	116	116	4月下旬	153	153
1月下旬	1668	1668	5月上旬	151	151
2月上旬	1402	1402	5月中旬	782	782
2月中旬	2449	2449	5月下旬	6816	6816
2月下旬	8954	8954	6月上旬	5159	5159
3月上旬	5529	5529	6月中旬	1700	1700
3月中旬	6114	6114	6月下旬	42	42
3月下旬	168	168	7月上旬	101	101
4月上旬	187	187			

本文と投稿日時の情報を得た。データの preprocessing として、ツイート本文の半角を全角に変換し、また、ツイート本文に書かれた数字を 0 に統一した。さらにツイート本文の改行とスペースを削除した。

区間は月ごとに上旬 (1 日～10 日)、中旬 (11 日～20 日)、下旬 (21 日～月末) と設定した。収集したツイートの投稿日時は 2020 年 1 月上旬から 7 月上旬であるため、設けた区間の総数は 19 となった。表 1 に、各区間のデータ数を示す。BERT の実装は自然言語処理ライブラリである Transformers の BertForSequenceClassification を利用して行った。

#### 4.1.2 結果と考察

19 個の区間のデータ数と Accuracy に対応する点をプロットした結果を図 3 に示す。横軸はデータ数、縦軸は Accuracy である。得られた 19 個の区間のデータ数と Accuracy をもとに、相関係数を求めた。全ての区間のデータ数と Accuracy の相関係数を求めると、相関係数は  $-0.45$  であった。また、データ数の少ない区間は Accuracy の振れ幅が大きいと考えたため、データ数が 1,000 未満の区間を除いて相関係数を求めた。相関係数は  $-0.69$  であった。

また、負の相関が大きいので、19 の点に対し、回帰直線を求めた。回帰係数は  $-3.5 \times 10^{-6}$ 、切片は 0.88 であった。回帰係数から、区間のデータ数がおおよそ 2,800 個増えるごとに、その区間の Accuracy は平均して 0.01 下がることがわかった。設定した 19 の区間のうち大部分の区間で、Accuracy は 0.85 から 0.95 の間に収まることを考えると、データ数が Accuracy に与える、影響は大きいと考えた。

相関係数と回帰係数から、ツイートの収集後、各区間でツイート数が大きく異なっている場合は、一時的な出来事が発生している区間の誤検出を減らすために、各区間のデータ数を統一する必要があると考えられる。

## 4.2 実験 2

分類精度が低い区間には、一時的な出来事が発生しているか調べる実験を行った。具体的には、提案手法の 3.1 節から 3.4 節までを行い、分類精度が低い区間には、一時的な出来事の発生件数が増える傾向にあるかどうかを確かめることによって行った。また、この実験では一時的な出来事の範囲を新型コロナウイルス関連の出来事に限定し、実験に用いるツイートや一

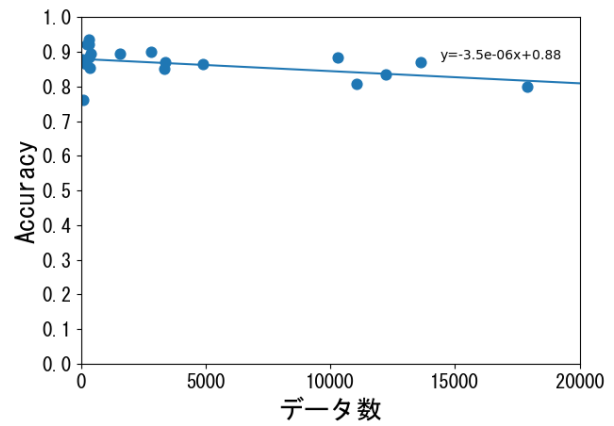


図 3 区間のデータ数と Accuracy の関係

時的な出来事の発生件数は新型コロナウイルスに関係するものを用いた。区間の設定方法は、二通り用意した。また、一時的な出来事の発生件数は、ニュース記事から一時的な出来事の件数を数え上げた。

#### 4.2.1 ラベル付きツイートと区間の設定

実験に用いたラベル付きツイートは、コロナウイルスに関係しているツイートが、事実であるか感想であるかラベル付けがされたツイートを用いた。データの収集として、はじめに COVID-19 日本語 Twitter データセットからコロナウイルスに関係するツイートで事実であるとラベル付けされたツイートの ID と、コロナウイルスに関係あるツイートで感想であるとラベル付けされたツイートの ID を取得した。その後、取得したツイートの ID からツイート本文と投稿日時の情報を得た。データの preprocessing として、ツイート本文の半角を全角に変換し、また、ツイート本文に書かれた数字を 0 に統一した。さらにツイート本文の改行とスペースを削除した。

また、区間の設定は、(a) 分割する区間の長さを長くする場合と (b) 分割する区間の長さを短くする場合の二通りを用意した。具体的には、(a) では、区間を月ごとに上旬 (1 日～10 日)、中旬 (11 日～20 日)、下旬 (21 日～月末) と設定した。また (b) では、上旬、中旬、下旬をさらに、それぞれ三つに分けて区間を設定した。

また、ツイートの収集後、各区間でデータ数を可能な限り統一した。各区間でデータ数を可能な限り統一する理由は、データ数が少ない区間ほど、分類精度は高くなる傾向にあり、それが一時的な出来事が発生する区間の誤検出につながると考えるからだ。具体的には、(a) の場合において、データ数が 2,000 を超える区間は、データを取り除いてデータ数を 2,000 にした。またデータ数が 2,000 未満の区間のうち、データ数が 200 を超える区間は、データを取り除いてデータ数を 200 にした。また、(b) の場合では、データ数が 660 を超える区間は、データを取り除いてデータ数を 660 にした。

なお、区間の設定が (a) の場合において、1 区間のデータ数を 200 と 2000 に統一した理由は、1 区間のデータ数を 200 で統一すると、取り除くデータ数が多くなり、モデルに学習させるデータが少なくなり過ぎるためである。

表 2 各区間のデータ数

区間	事実	感想	区間	事実	感想	区間	事実	感想	区間	事実	感想
1月1日～10日	25	25	1月1日～3日	0	0	1月4日～6日	0	0	1月7日～10日	25	25
1月11日～20日	100	100	1月11日～13日	9	9	1月14日～16日	57	57	1月17日～20日	49	49
1月21日～31日	1000	1000	1月21日～23日	168	168	1月24日～26日	330	330	1月27日～31日	330	330
2月1日～10日	1000	1000	2月1日～3日	330	330	2月4日～6日	330	330	2月7日～10日	330	330
2月11日～20日	1000	1000	2月11日～13日	330	330	2月14日～16日	330	330	2月17日～20日	330	330
2月21日～29日	1000	1000	2月21日～23日	330	330	2月24日～26日	330	330	2月27日～29日	330	330
3月1日～10日	1000	1000	3月1日～3日	330	330	3月4日～6日	330	330	3月7日～10日	330	330
3月11日～20日	1000	1000	3月11日～13日	330	330	3月14日～16日	330	330	3月17日～20日	330	330
3月21日～31日	100	100	3月21日～23日	24	24	3月24日～26日	37	37	3月27日～31日	96	96
4月1日～10日	100	100	4月1日～3日	44	44	4月4日～6日	68	68	4月7日～10日	75	75
4月11日～20日	100	100	4月11日～13日	73	73	4月14日～16日	52	52	4月17日～20日	72	72
4月21日～30日	100	100	4月21日～23日	59	59	4月24日～26日	49	49	4月27日～30日	45	45
5月1日～10日	100	100	5月1日～3日	37	37	5月4日～6日	57	57	5月7日～10日	57	57
5月11日～20日	100	100	5月11日～13日	43	43	5月14日～16日	54	54	5月17日～20日	330	330
5月21日～31日	1000	1000	5月21日～23日	330	330	5月24日～26日	330	330	5月27日～31日	330	330
6月1日～10日	1000	1000	6月1日～3日	330	330	6月4日～6日	330	330	6月7日～10日	330	330
6月11日～20日	1000	1000	6月11日～13日	330	330	6月14日～16日	316	316	6月17日～20日	35	35
6月21日～30日	42	42	6月21日～23日	27	27	6月24日～26日	0	0	6月27日～30日	15	15
7月1日～10日	100	100	7月1日～3日	49	49	7月4日～6日	52	52	7月7日～10日	0	0

収集したツイートの投稿日時は 2020 年 1 月上旬から 7 月上旬であるため、(a) では設けた区間の総数は 19 となった。また、(b) では設けた区間の総数は 57 となった。なお、これらの区間の総数には、収集したツイートが少なく、データ数を統一できなかった区間も含まれている。また、区間の設定が (b) の場合において、データ数を統一できなかった区間の分類精度の算出は行わなかった。表 2 に (a) 及び (b) の各区間のデータ数を示す。

#### 4.2.2 ニュース記事から一時的な出来事の件数を数え上げるための手順

ニュース記事から新型コロナウイルスに関する一時的な出来事の件数を数え上げるために、NHK の新型コロナウイルス特設サイトの時系列ニュース<sup>4</sup>を利用した。新型コロナウイルス特設サイトの時系列ニュースでは、新型コロナウイルスに関連した記事を時系列順にまとめて公開している。

はじめに、第一著者がその Web ページに記載されている記事を 1 つずつ読み、記事 1 つにつき記事内の出来事が一時的な出来事であるかそうでないかを判定した。なお、時系列ニュースの記事には、記事 1 つにつき複数の出来事を紹介している場合もある。その場合は最も主要な出来事を判定の対象とした。一時的な出来事であるか、そうでないかの判定基準は、出来事が発生し、終了するまでの期間が二週間未満であれば一時的な出来事、二週間を超えるのであれば一時的な出来事ではないとした。また、繰り返し発生している出来事も一時的な出来事ではないと判定する。従って例えば、一日の感染者数（陽性者数）を紹介している記事内の出来事は、繰り返し発生しているため、一時的な出来事ではないと判定した。また、出来事が発生し、終了するまでの期間が特定できない出来事は判定を行わなかった。

記事内で紹介されている出来事が一時的な出来事であるかど

うかの判定を行った後、一時的な出来事が、設定した区間で何件発生しているのかを数え上げた。なお、一時的な出来事が発生している期間が 2 つの区間に渡って存在する場合は、出来事が発生している期間が最も長く含んでいる区間にあるものとみなした。また、一時的な出来事が発生している期間が完全に 1 つあるいは複数の区間を含んでいる場合は、それらの区間全てで、その出来事が発生しているとみなした。

#### 4.2.3 結果と考察

一時的な出来事の発生件数と Accuracy に対応する点をプロットした。区間のデータ数の違いが分類精度に影響を与えないようにするため、区間の設定が (a) の場合においては、データ数が 200 及び 2,000 ある区間に限定してプロットを行った。また、区間の設定が (b) の場合では、データ数が 660 ある区間に限定してプロットを行った。それらの結果を図 4 にそれぞれ示す。横軸は一時的な出来事の発生件数、縦軸は Accuracy である。

プロットした点において、回帰係数を求めたところ、(a) については、データ数が 200 ある区間の回帰係数は  $-0.0022$ 、データ数が 2,000 ある区間の回帰係数は  $-0.0034$  であった。また (b) については、データ数が 660 ある区間の回帰係数は  $-0.021$  であった。実験の結果から、(a) の場合では、分類精度と一時的な出来事の発生件数は関係がないという結果が得られた。一方、(b) の場合では、分類精度が低い区間には、一時的な出来事の発生件数が増加する傾向にあるという結果が得られた。

また、分類器の判定が正解と異なっているツイートの特徴について調べることを行った。分類器の判定と正解が異なったツイートには、表 3 に示すように、一時的な出来事に関するツイートが見つかった。しかし、二週間以上続いた出来事について言及するツイート、つまり、今回では一時的ではないと判定されるべき出来事のツイートも見つかった。例えば、表 4 に示すように、全国各地のテーマパークが 3 月頃から臨時休園を行

4: <https://www3.nhk.or.jp/news/special/coronavirus/chronology/>



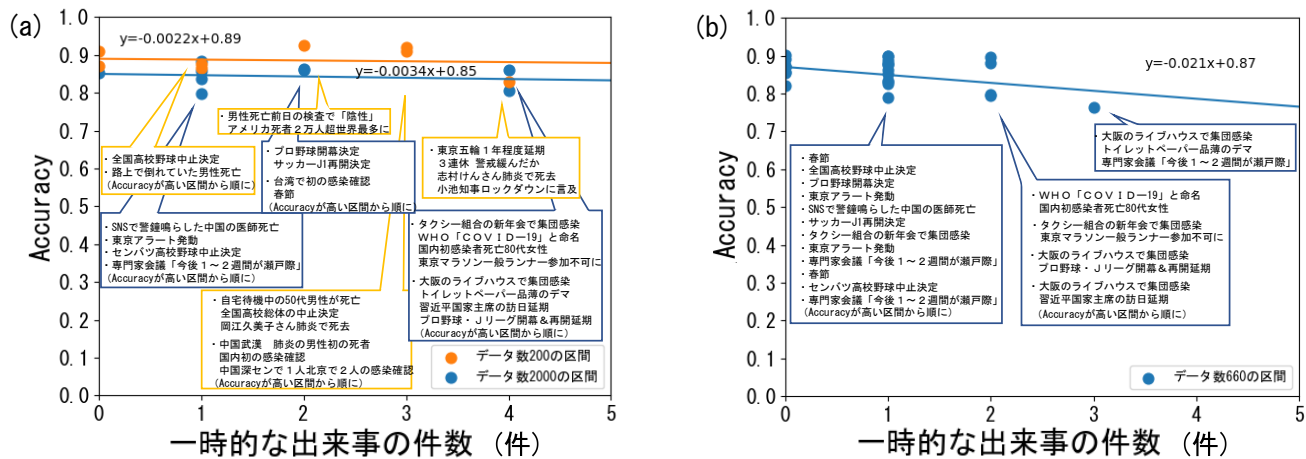


図 4 一時的な出来事の発生件数と Accuracy の関係

う出来事について述べたツイートや、3月2日から春休みの期間まで、全国の小学校と中学校、及び高校が臨時休校となった出来事に関するツイートである。本稿の実験では、繰り返し発生していない出来事かつ、出来事が続いた期間が二週間未満の出来事を一時的な出来事と判定したが、今回の実験の結果をうけて、今後どのくらいの期間の長さが一時的といえるか、区間の長さとともに、検討する必要があることが分かった。

また、実験に用いたツイートには、ニュースには取り上げられないような小規模な出来事についてのツイートも含まれていた。そのため、分類器の判定が正解と異なっているツイートには、表5のように、ニュースには取り上げられないような小規模な出来事かつ、一時的な出来事に関して書かれたツイートも多数見つかった。一時的かつ、ニュースに取り上げられるような出来事が発生した区間を検出する場合、事前にツイートから小規模な出来事に関して書かれたツイートを取り除くなどの処理が有効であると考えられる。

## 5 おわりに

本研究では、社会での変化や出来事の発生を、ソーシャルメディアのデータを基に自動的に提示するということを目標とした。そして、本稿では、一時的な出来事を定義し、一時的とは無関係なラベル付けがされたツイートと、機械学習を用いて、一時的な出来事の発生の検出方法を提案した。提案手法では、はじめに、ツイートの投稿日時を区間に分割した。そして、区

間ごとに、学習用データと評価用データを定め、それらのデータを用いて、分類器の構築と区間の分類精度の算出を行った。ある区間の学習用データは、その区間に属していないツイートとした。また、評価用データは、その区間に属しているツイートとした。一時的な出来事の発生の検出は、区間単位で行い、分類精度が低い区間で、一時的な出来事が発生しているとした。本稿では二つの実験を行った。

一つ目の実験では、各区間のデータ数のばらつきが、分類精度にどのくらい影響を与えているか調べた。実験結果からは、各区間のデータ数のばらつきが分類精度に影響を与えていることが得られた。具体的には、実験で用意した全ての区間において、区間のデータ数と Accuracy の相関係数は  $-0.45$  であった。この結果から、一時的な出来事が発生した区間の誤検出を減らすために、各区間のデータ数を統一する必要があることが分かった。

二つ目の実験では、提案手法の区間の分類精度と一時的な出来事の発生件数には相関があるか調べた。また、この実験では、区間の設定方法は、(a) 分割する区間の長さを長くする場合と、(b) 分割する区間の長さを短くする場合の二通りを用意した。実験結果として (a) では、分類精度と一時的な出来事の発生件数に関係がないという結果が得られた。しかし、(b) では、分類精度が低い区間ほど、一時的な出来事の発生件数は増加していた。具体的には、一時的な出来事が一件発生するごとに、区間の分類精度は  $0.021$  下がった。

表 3 一時的な出来事に関するツイート

正解	分類器	ツイート
感想	事実	韓国の病院で精神科病棟の入院患者 102 人のうち 99 人が新型コロナ感染、中国ネットにも衝撃 この事例はもっと日本で知らるべきである。
感想	事実	中国からの飛行機がガンガン飛んでる時点でパンデミックは見えてるやろww【速報】都内のタクシー運転手、コロナ感染
事実	感想	コロナったタクシー運転手の会社すぐ近所やんけ www
事実	感想	コロナコロナるせえな！トイレトペーパー売り切れてどこにも売ってねえよ！
感想	事実	新型コロナ座談会 連鎖断てるか、この 1~2 週が正念場:日本経済新聞 良い記事を読めた。

表 4 2 週間以上続いた出来事に関するツイート

正解	分類器	ツイート
事実	感想	ディズニーランドが明日から休園だって！ #新型コロナウイルス
感想	事実	ディズニーシー&ランド臨時休園！サンリオピューロランドに続いて休園…休園は 2 月 29 日から 3 月 15 日まで旅行の計画してる方見直し必須ですね詳細は東京ディズニーリゾートのオフィシャルサイトをご確認ください#TDL #TDS#コロナウイルス対策
事実	感想	臨時休校は本当に意味のない新型コロナ対策なのか？ - Y!ニュース
事実	感想	こちらがコロナウイルス対策が瀬戸際で政府が休校や不要不急の外出を避けるよう要請している最中の朝の品川駅です。ご査収下さい。いやまあ、地面が見えてるだけいつもよりだいぶ空いてますけども。あと近所の学校は普通に登校してましたね。

表 5 小規模かつ一時的な出来事に関するツイート

正解	分類器	ツイート
感想	事実	卒業式がなくなりました。私の大学生活が突然終わりました。数ヶ月前から着ると決めて予約していた袴を卒業式で着れなくなりました。私よりお母さんがブチギレています。一気に家庭の雰囲気暗くなりました。コロナが憎い。
事実	感想	今日は中学校の卒業式でした。最後は歌を歌って終わり。コロナなんて忘れちゃうくらい盛り上がりでした。最後の 2 週間を過ごさせてやりたかった。本当にいいクラスでした。
事実	感想	コロナで従姉妹の結婚式が延期になってしまった…
感想	事実	コロナの影響で、お仕事なくなったため、しばらくお弁当はお休み。おうちで夜ご飯の残りを食します。。

実験により、提案手法の区間の分類精度と、一時的な出来事の発生件数には相関がある可能性があることを示した。これにより、一時的とは無関係なラベル付けがされたツイートをを用いて、自動的に一時的な出来事が発生した区間の提示ができる可能性があることが分かった。しかし、分割する区間の長さを変えることで、検出可能な一時的な出来事がどのように変化するかについては、本稿では未だ説明できていない。そのため、今後は分割する区間の長さ、検出可能な出来事との関係を詳しく調べることが必要である。また、本稿では、各区間の分類精度を算出した後において、分類精度が低い区間を自動的に検出するための方法が提示できていない。従って今後、一時的な出来事が発生した区間を自動的に提示するために、統計的手法を用いて分類精度が低い区間を見つける方法を検討することが必要である。

**謝辞** 本研究の一部は JSPS 科研費 18H03342, 19H04221, 19H04218, および大川情報通信基金の助成を受けたものです。

## 文 献

- [1] 羽山徹彩. Twitter データを用いたテレビ番組ダイジェスト自動生成に対するユーザ分類の適用. Technical Report 14, 金沢工業大学, 2014.
- [2] Jon Kleinberg. Bursty and hierarchical structure in streams. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Vol. 7, pp. 91–101, 2002.
- [3] 坂本翼, 廣田雅春, 横山昌平, 福田直樹, 石川博. Twitter ストリームのバーストの断続性に着目したキーワード抽出. 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2012), C7-3, 2012.
- [4] 蝦名亮平, 中村健二, 小柳滋. リアルタイムバースト解析手法の提案. 情報処理学会論文誌データベース (TOD), Vol. 5, No. 3, pp. 86–96, 2012.
- [5] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo.

Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, p. 851–860, New York, NY, USA, 2010. Association for Computing Machinery.

- [6] 水沼友宏, 池内淳, 山本修平, 山口裕太郎, 佐藤哲司, 島田諭. Twitter におけるバーストの生起要因と類型化に関する分析. 情報社会学会誌, Vol. 7, No. 2, pp. 41–50, 2013.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding, 2018.