発生規模と時系列を考慮した Twitterイベントにおける偽情報の早期自動検出

山中　仁斗† 　張　 建偉†

† 岩手大学理工学部　〒 020–8551 岩手県盛岡市上田 4-3-5
E-mail: †{s0616062,zhang}@iwate-u.ac.jp

あらまし　近年の SNS の発達において, 偽情報の拡散が問題となっている. 偽情報は, 政治・経済・災害等の面で世間に悪影響を及ぼす場合があるが, 人手での素早い検出は難しく, 自動で早期検出を行う技術の開発が求められている. 本研究では, Twitter を対象に, イベントの発生規模とイベントを構成するツイートの時系列を考慮した, 機械学習による 2STEP の偽情報の早期検出手法を提案する. STEP1 では, イベントの発生初期に真偽の決定を試み, 偽情報の早期検出を図ると共に, 真偽が決定された一部のイベントを追跡対象から取り除くことで検出効率の向上を目指す. STEP2 では, STEP1 で真偽を判別できなかったイベントを追跡対象とし, 時系列に沿って真偽を判別していく. 5 つの機械学習モデルを使用して実験を行った結果, STEP1, STEP2 ともに SVM が最適なモデルであり, 本手法を用いることで早期検出を実現できるという結果が得られた.

キーワード　Twitter，時系列データ, 機械学習, 偽情報, 早期検出

1 はじめに

近年, パソコンやスマートフォンの普及に伴い, ソーシャルネットワーキングサービス（以下, SNS）の利用者が増加してきた. SNS は有益な情報源として活用できる反面, 事実とは誤った情報（以下, 偽情報）がユーザーへ拡散され, 世間に悪影響を及ぼす場合がある. 例えば, 2013 年に, 「米国・ホワイトハウスが爆発しバラク・オバマ氏が負傷した」という偽情報が 130 億ドルの株価暴落を引き起こした [1]. 国内では, 災害発生時に偽情報が拡散される事例が多く挙げられ [2], 2018 年には豪雨被災地で「レスキュー隊に変装した窃盗グループが潜んでいる」といった偽情報が拡散され混乱を生み出した [3].

人手では, SNS 上の無数のデータを追跡することは難しく, 情報が誤っていると分かる頃には既に拡散済みである場合が多い. そのため, 偽情報を自動的に, より早く検出する技術が求められており, 数々の研究が行われてきた [4, 5, 6, 7, 8, 9, 10].

SNS では世界中のユーザーによって絶え間なく情報の投稿が行われているが, 投稿の発生規模を考慮している先行研究は我々が探した限り見つかっていない. また, SNS 上のデータは時間の経過につれて蓄積していくが, 先行研究の多くはデータを静的なまとまりとして捉え, 最終的に得られるデータ（蓄積済みの全てのデータ）から特徴量を作成しているため, 早期での検出に対応できていない.

我々は, SNS の一つである Twitter を対象とし, データの発生規模と時系列を考慮した, 機械学習による偽情報の早期検出手法を提案する. 我々の検出手法は 2 段階のフェーズに分けられる（以下, STEP1・STEP2）. STEP1 では, イベントの発生初期に真偽の決定を試み, 偽情報の早期検出を図ると共に, 真偽が決定された一部のイベントを追跡対象から取り除くことで, 検出効率の向上を目指す. STEP2 では, STEP1 で真偽を判別できなかったイベントを追跡対象とし, 時系列に沿って真偽を判別していく. 5 つの機械学習モデルを使用して実験を行った結果, STEP1, STEP2 ともに SVM が最適なモデルであり, 本手法を用いることで早期検出を実現できるという結果が得られた.

2 関連研究

Srijan ら [4] や Zhou ら [5], Lillie ら [6] は, 偽情報が世間に及ぼす影響や, 拡散に関与するユーザー, 検出アルゴリズム等に関して, サーベイ論文をまとめている.

Castillo ら [7] や Qazvinian ら [8] は Twitter において, Yang [9] らは Sina Weibo において, ユーザーの情報や, メッセージの内容および拡散パターンから多数の特徴量を作成し, 偽情報の検出を試みた. これらの先行研究では, SNS から得られたデータを始めから全て利用できるという前提で実験を行っており, 時間の経過によるデータの変動を考慮していないため, 早期検出には対応できていないという課題がある.

Ma ら [10] は Twitter において, データの時系列を考慮し, 一定時間ごとに特徴量の変化量を求め, その変化量を特徴量として使用することで, 偽情報の早期検出を行う手法を提案した. 当研究では, 時間の経過が考慮されていないという先行研究の課題を解決できているが, SNS におけるデータの発生規模が考慮されていないという課題が残っている.

本研究では, 情報の発生規模と時系列の両方を考慮した, 2 段階の検出手法を提案する点でこれらの研究とは異なっている.

3 提案手法の概要

偽情報の検出はイベント単位で行うものとする. イベントとは「○○氏が亡くなった」といった特定の出来事を指し，1 つ

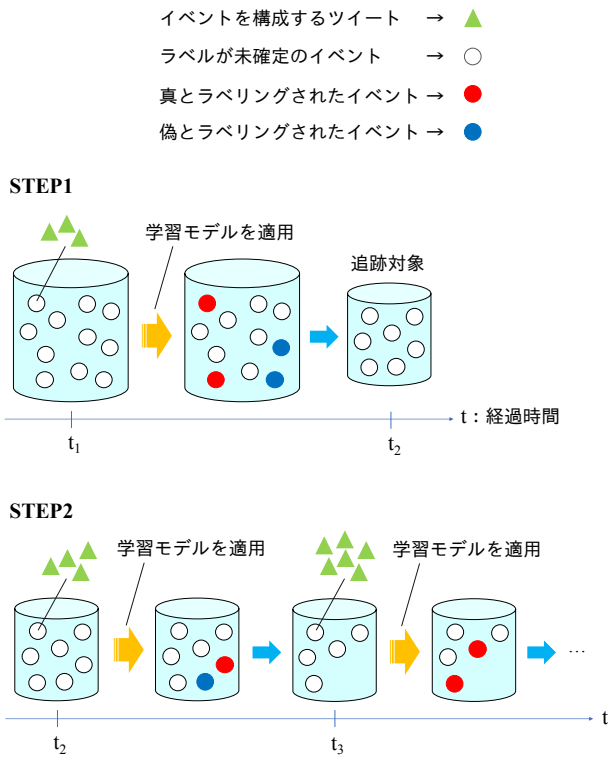

図 1　提案手法の全体像

のイベントは多数のツイート [1]から構成される．

提案手法の全体像を図 1 に示す．本手法おける検出過程は，イベントの発生初期に，一部のイベントに真偽のラベルを付与する STEP1 と，ラベルを付与されていない残りのイベントに対して，時系列に沿って真偽を判別していく STEP2 に分けられる．

STEP1

イベントの発生規模を考慮する際，全てのイベントを追跡するのは検出効率が悪いため，イベント発生の初期段階で，確実に真または偽と判断できるものを，このフェーズであらかじめ取り除く．

初期段階で使用できるツイートから，イベント単位で特徴量を作成する．その後，最適な機械学習モデルを利用して各イベントの真偽の確率を算出し，予測確率の高いイベントのみに真偽ラベルを付与することで実現する．

一部のイベントについて初期段階で真偽を判別することで，偽情報の早期検出を図ると共に，追跡対象の数を減らすことにより，検出効率の向上を図る．

STEP2

STEP1 で決定した追跡対象に対して，時系列に沿って偽情報の検出を行う．各イベントのツイート集合に，新しく投稿されたツイートを追加しながら，一定時間ごとにイベント単位で特徴量を作成し，最適な機械学習モデルによって偽情報の検出を繰り返す．

1 つのイベントに対するツイートは時間の経過に連れて増加していくため，作成できる特徴量も時間によって異なる．すなわち，時間によって真偽の予測結果が異なる場合があると考えられるため，STEP2 では，イベントに付与する真偽のラベルをどの時点で確定させるか，明確な条件を設ける必要がある．我々は，

- 真偽の予測確率が高くなった場合，その予測は信用できる．
- 真偽の予測結果が一定以上の時間不変ならば，その予測は信用できる．

という仮定から，STEP1 と合わせて図 2 のような検出フローを提案する．図 2 において，t はイベント発生後の経過時間，$P_P(e_i)$, $P_N(e_i)$ はイベント e_i に対する偽，真それぞれの予測確率，τ は検出の基準とする予測確率の閾値，n は検出処理の繰り返し回数を表す．以下，本研究では図 2 と同様に，偽のイベントを P（Positive：陽性），真のイベントを N（Negative：陰性）と表現する．

1：Twitter 上でユーザーが投稿するメッセージを指す．

表 1　使用したデータセットの詳細

統計値	イベントの収束時間（時間）	イベントに含まれるツイート数（個）
最大値	507.784	8084
最小値	1.295	10
平均値	35.633	320.043
中央値	31.49	54

4　実験準備

4.1　データセット

Castillo ら [7] によって作成されたデータセットを使用した．彼らは，2010 年 4 月から 9 月にかけて，Twitter Monitor [11] を使用することで，288 件のイベントと，各イベントを構成するツイート集合を抽出した．本研究では，その中でツイート数が 10 未満のイベントを取り除き，真のイベント 112 件，偽のイベント 99 件，合計 211 件のイベントを用いた．使用したデータセットの詳細を表 1 に示す．表 1 において，イベントの収束時間とは，イベントを構成しているツイート集合の中で，最も早い日時に投稿されたツイートから，最も遅い日時に投稿されたツイートまでの経過時間を表す．

4.2　特徴量の作成

Castillo ら [7] や Ma ら [10] の研究を参考に，偽情報の検出に効果的であると考えられる 52 次元の特徴量を作成した．これらの特徴量はメッセージベース，ユーザーベース，拡散ベースの 3 つのカテゴリに分類することができる．なお，メッセージベースの特徴量は英文データのみを用いて作成した．

特徴量は，イベントを構成するツイート集合から得られるデータを，イベント単位で集約することで作成する．例えば，「ツイートに URL が含まれているかどうか」といったツイート単位の bool 値データは，「URL が含まれているツイートの割合」というイベント単位の特徴量に変換され，「ツイートの文字数」といったツイート単位の数値データは，その平均値や最大値が特徴量として用いられる．

作成した特徴量を表 2 に示す．いくつかの特徴量に関して以下で言及する．

表 2　特徴量一覧

カテゴリ	特徴量
メッセージベース 29次元	ツイートの大文字の割合
	大文字の割合が30%を超えるツイートの割合
	ツイートがWEBアプリから投稿された割合
	ツイートがスマートフォンから投稿された割合
	ツイートがサードパーティアプリから投稿された割合
	！を含むツイートの割合
	？を含むツイートの割合
	？や！を複数含むツイートの割合
	メンション（@）を含むツイートの割合
	ハッシュタグ（#）を含むツイートの割合
	ネガティブな顔文字を含むツイートの割合
	各ツイートのネガティブな単語の個数の平均値
	ポジティブな顔文字を含むツイートの割合
	各ツイートのポジティブな単語の個数の平均値
	各ツイートの感情スコアの平均値
	一人称代名詞を含むツイートの割合
	二人称代名詞を含むツイートの割合
	三人称代名詞を含むツイートの割合
	各ツイートの文字数の平均値
	各ツイートの単語の個数の平均値
	各ツイートのURLの個数の平均値
	URLを含むツイートの割合
	月～日曜日に投稿されたツイートの割合（7次元）
ユーザーベース 9次元	プロフィールの説明文を記入しているユーザーの割合
	プロフィール画像を設定しているユーザーの割合
	プロフィールに場所を設定しているユーザーの割合
	プロフィールにURLを設定しているユーザーの割合
	ユーザーがアカウントを作成してからの経過秒数
	ユーザーのフォロワー数の平均値
	ユーザーのフォロー数の平均値
	ユーザーの合計ツイート数の平均値
	公式ユーザーの割合
拡散ベース 14次元	ツイートの数
	いいねの数の平均値
	いいねの最大数
	リツイート数の平均値
	リツイートの最大数
	全ツイートの内リツイートである割合
	ルートの数
	最大の木のノード数
	各ノードの深さの平均値
	各ノードの深さの最大値
	ルートの次数の平均値
	ルートの次数の最大値
	ルート以外のノードの次数の平均値
	ルート以外のノードの次数の最大値

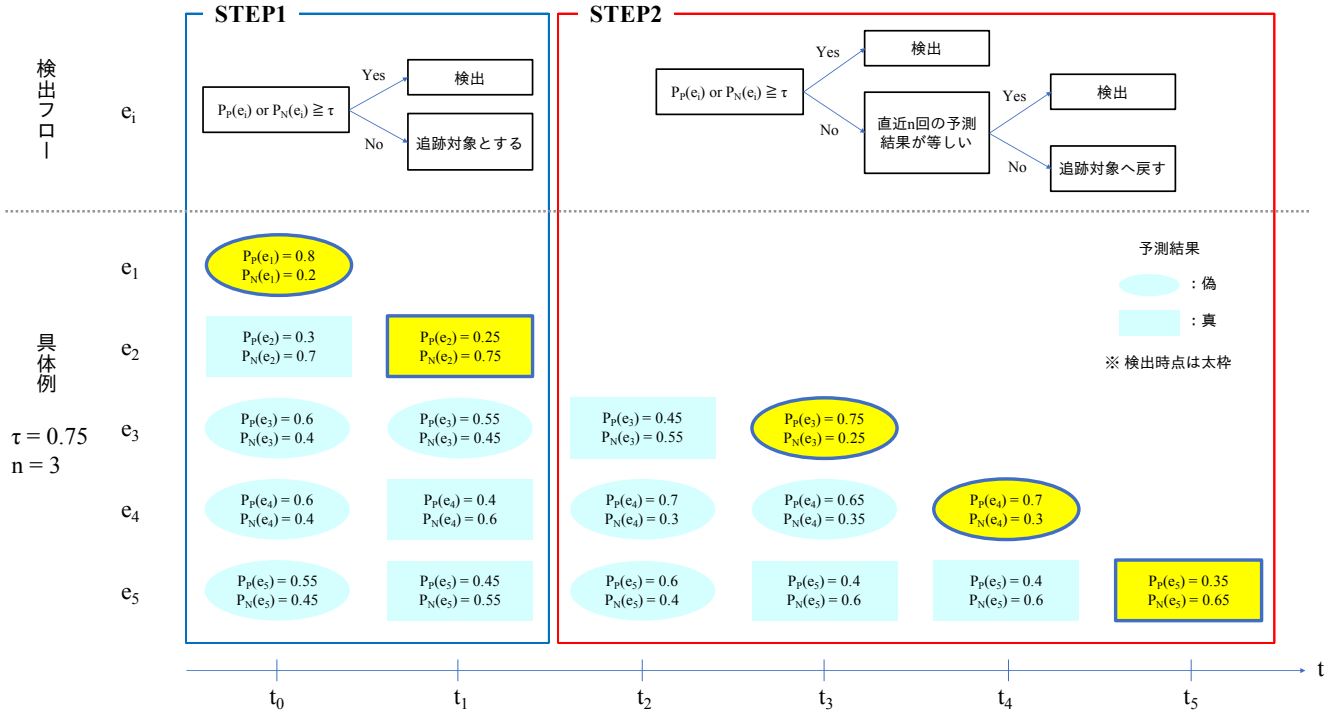

図 2　検出フローと具体例

感情スコア

感情スコアを算出するために, 感情分析ツールの VADAR [12] を用いた. VADER とは, 辞書とルールの組み合わせによって感情値を求めるパッケージである. 7,516 語のエントリを持つ辞書から単語感情値を求め, その値を疑問符や感嘆符, 強調語, 否定語などに関するルールによって修正することで, 入力した文章の感情値を算出できる.

本研究では, 各ツイートに対して, VADER における総合的な感情の評価値として定義されている compound 値を求め, その値を対象ツイートの感情スコアとして利用した.

拡散ベースの特徴量

表 1 に示した拡散ベースの特徴量において, 下 8 つの特徴量（ルートの数～ルート以外の次数の最大）は, リツイートのネットワーク構造から木を作成することで算出した. 木の例を図 3 に示す. 図 3 において, 深さとは, あるノードに対して, ルートに至るまでの経路数（辺の数）を表す. 次数とは, ノードに接続している辺の数を表す. 例えば図 3 において, 深さが最大であるのは ”ユーザー F” のノードであり, 次数が最大であるのは ”ユーザー H” のノードである.

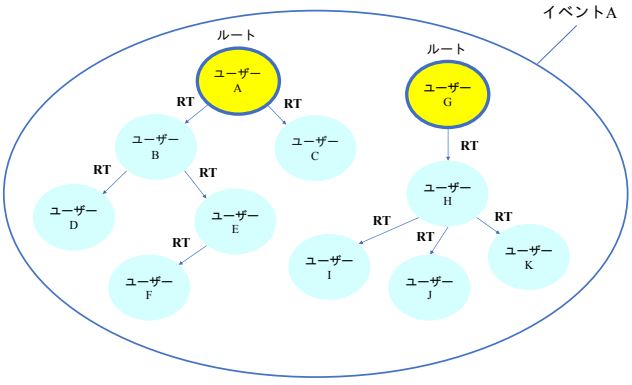

図 3　木　の　例

4.3　機械学習モデルの作成

STEP1, STEP2 のそれぞれにおいて, どの機械学習モデルの使用が適切であるかを検証する. 今回比較したモデルは, ロジスティック回帰 (LR), 非線形サポートベクタマシン (SVM), ランダムフォレスト (RFC), 勾配ブースティングマシン (GBC), 多層パーセプトロン (MLP) の 5 つである. 各モデルにおいて, 学習データを用いてグリッドサーチによるパラメータチューニングを行い, 最適なパラメータを設定してモデルを作成した.

5　評 価 実 験

5.1　データセットの使用方法

データセットを 5 分割し, 4/5 を学習データ, 1/5 をテストデータとして用いる. 学習データでモデルを作成し, テストデータで評価を行う. テストデータの特徴量は, イベント発生から何時間分のツイートを使用するかによって変化する. 実験を行う際には, 学習データとテストデータを変更しながら結果を合計 5 回算出し, 5 回の平均値を最終的な評価値とする.

5.2　STEP1 のモデル検証

まず, STEP1 において最適な機械学習モデルの検証を行う. 本 STEP の目的は, イベント発生の初期段階で, 真偽の可能性が高いと判断できるイベントを検出することで, 偽情報の早期検出を図ると共に, 追跡対象の数を減らすことである. 各イベ

t = 5 の予測結果

テストデータ

順位	偽の確率	真の確率	予測結果	実際の結果	正誤
1	**0.9**	0.1	偽	偽	正解
2	**0.8**	0.2	偽	偽	正解
3	**0.7**	0.3	偽	真	不正解
4	**0.55**	0.45	偽	偽	正解
3	0.25	**0.75**	真	偽	不正解
2	0.2	**0.8**	真	真	正解
1	0.1	**0.9**	真	真	正解

$AP_P^{(5)} = \frac{\frac{1}{1} + \frac{2}{2} + \frac{3}{4}}{3}$

$AP_N^{(5)} = \frac{\frac{1}{1} + \frac{2}{2}}{2}$

図 4　平均適合率の算出例

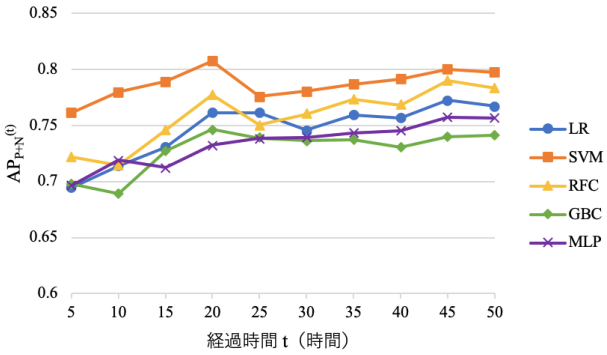

図 5　各モデル間の $AP_{P+N}^{(t)}$ の比較

ントに対して真偽の予測確率を算出し, 初期段階において, 高い確率を示したイベントがどれだけ正しくラベル付けされているかという観点で各モデルの評価を行う.

5.2.1　評 価 指 標

イベント発生から t 時間後における, 各イベントの真偽の予測確率に対して, 真偽それぞれの予測確率が高い結果から, 予測確率が最も 0.5 に近い結果を最下位とした順位付けを行い, 平均適合率 (AP:Average Precision) を算出する. 平均適合率は式 (1) で定義され, 順位のある予測結果に対して, 正解データがより上位に集まるほど高い値を示す指標である. P(k) は上位 k 番目までの結果を用いた適合率, f(k) は k 番目の出力が正解であれば 1, 不正解であれば 0 を出力する関数を表す.

$$AP = \frac{\sum_{k=1} P(k) \times f(k)}{(\text{正解データの数})} \tag{1}$$

t 時間後における偽のイベントと真のイベントを対象とした平均適合率を, それぞれ $AP_P^{(t)}$, $AP_N^{(t)}$ とすると, 平均適合率の算出例は図 4 のようになる. 図 4 では, イベント発生から 5 時間後における平均適合率を求めている.

$AP_P^{(t)}$ と $AP_N^{(t)}$ の平均値を $AP_{P+N}^{(t)}$ とすると, 本 STEP に用いる評価指標 AP_{w_sum} は式 (2) で求められる.

$$AP_{w\_sum} = \sum_{t=kn} \gamma^{(t-k)} AP_{P+N}^{(t)} \tag{2}$$

ここで, $\gamma^{(t-k)} (0 < \gamma \leqq 1)$ は経過時間による重み, k（単位：時間）は経過時間の刻み幅, n は自然数を表す. 本研究では, $\gamma = 0.5$, $k = 5$, $n = 1 \sim 10$ に設定して実験を行った.

5.2.2　実 験 結 果

実験結果を図 5, 表 3 に示す. 図 5 より初期段階から SVM が優れた結果を示していることが読み取れ, 表 3 より SVM が総合的に最も高い値を示していることが分かる.

SVM には, データがわずかな特徴量しか持たない場合にも複雑な決定境界を生成できるという性質があるため, このような結果を示したと考えられる.

ここで, SVM を用いたモデルによって, イベント発生の初期段階で, 予測確率の高いイベントが正確にラベル付けされていることを実際に確認する. 初期段階における, 真偽それぞれのイベントを対象とした適合率を, 予測確率ごとに算出した結果を表 4 に示す. 表 4 において, t はイベント発生後の経過時間,

表 3　各モデルの AP_{w_sum}

学習モデル	LR	**SVM**	RFC	GBC	MLP
AP_{w_sum}	0.718	**0.787**	0.745	0.72	0.72

表 4　SVM による初期段階かつ予測確率の高いイベントの適合率

	P（偽）			**N**（真）		
τ \ t	5	10	15	5	10	15
0.85	1	null	null	1	1	1
0.8	1	1	1	0.912	0.912	0.833
0.75	0.833	0.833	0.667	0.971	0.893	0.875
0.7	0.566	0.683	0.608	0.748	0.745	0.795

τ は予測確率を表す. 偽のイベントに関して, t = 10, 15 かつ τ = 0.85 のとき, 偽であると予測されたイベントが存在せず適合率を計算できなかったため, null と表記した.

偽のイベント, 真のイベント共に, t が小さく, かつ τ が大きいほど良い結果を示す傾向があり, 実際に SVM が STEP1 に適していることが分かる.

5.3　STEP2 のモデル検証

STEP2 では, STEP1 で検出されなかったイベントを追跡対象とし, 時系列に沿って真偽を判別していく. 偽情報をより多く, 正確に検出することが最大の目的である. 本 STEP における最適なモデルを検証するため, イベント発生後の経過時間ごとに, 偽情報を対象として, 各モデルの F 値を算出し比較を行った. 本研究では, 経過時間 t = 5, 10, 15, ... , 50, all（単位：時間）として実験を行った. t = all は, イベントを構成する全てのツイートを使用した場合を指す. 結果を図 6 に示す. 各時点において SVM が高い F 値を示しており, STEP2 においても SVM が最適なモデルであることが分かった.

5.4　早期検出の検証

本提案手法を用いることで, 早期検出を実現できるかどうか検証を行う. 5.2 節, 5.3 節の結果より, STEP1, STEP2 で使用する学習モデルは SVM とする. 本研究では, イベント発生後の経過時間が 10 時間までの場合を初期段階であると想定し, 経過時間 t ≦ 10（単位：時間）のときは STEP1 を, t ＞ 10 のときは STEP2 の処理を行う. また, 図 2 の検出フローにおいて, 本研究では τ = 0.75, n = 3 として実験を行った.

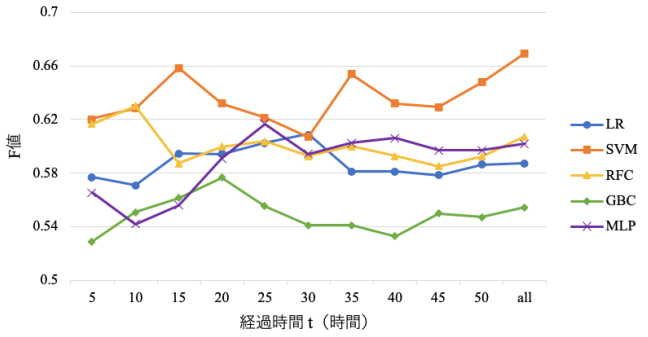

図 6　各モデル間の F 値の比較

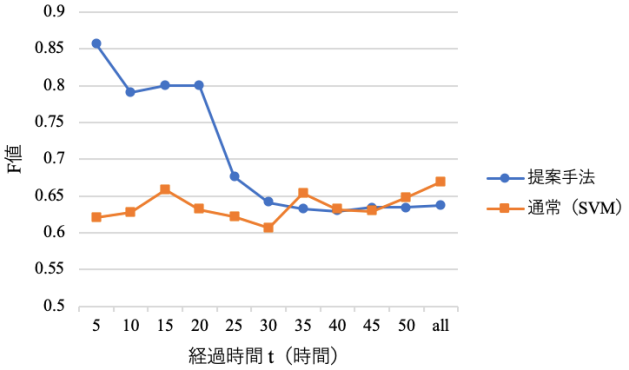

図 7　提案手法と通常の SVM による F 値の比較

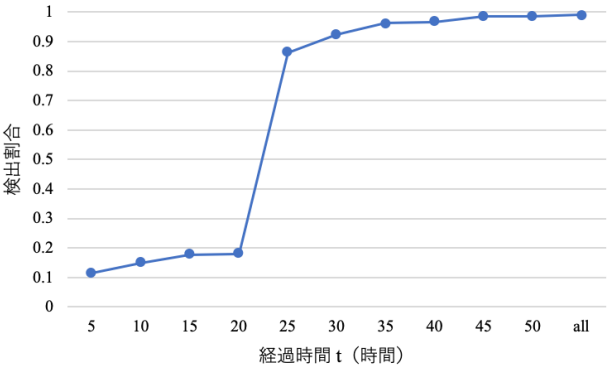

図 8　検出済みデータの割合

5.4.1　評 価 方 法

図 2 に示した提案手法を用いて, イベント発生後の経過時間ごとに, 偽情報を対象とした F 値を算出する. イベント発生後の経過時間ごとに, 提案手法を用いずに全てのイベントに対して SVM を適用した結果を比較対象とし, 提案手法を用いた場合の結果と比較する. 本研究では, 5.3 節と同様に, 経過時間 t = 5, 10, 15, ... , 50, all（単位：時間）とする. また, 各経過時間において, 全てのイベント数に対して, どれだけの割合のイベントを検出できているか確認する.

5.4.2　実 験 結 果

提案手法と通常の SVM による F 値の算出結果を図 7 に, 検出済みデータの割合を図 8 に示す. 図 7 より, t = 5〜30 の間は, 提案手法の F 値が通常の検出手法を上回っており, t = 30 を超えてからは同程度の F 値を示していることが読み取れる. 提案手法は通常の検出方法に比べ, より短い経過時間で高い F 値を示していることから, 偽情報の早期検出を実現できていることが分かる.

また, 図 8 より, 検出率は t = 20 までは全体の 20%に満たないが, t = 20〜25 にかけて大きく上昇し, t = 35 の時点では約 100%のイベントを検出できていることが読み取れる. 提案手法を用いた場合, ほとんど全てのイベントが, 発生から 35 時間で検出されることが分かった.

6　ま　と　め

本研究ではイベントの発生規模とツイートの時系列を考慮した, 機械学習による 2STEP の偽情報検出手法を提案し, STEP1, STEP2 のそれぞれにおける最適な機械学習モデルの検証, および提案手法が偽情報の早期検出を実現できるかどうかの検証を行った.

5 つの機械学習モデルを使用して実験を行った結果, STEP1, STEP2 ともに SVM が最適なモデルであり, 提案手法を用いることで, 偽情報の早期検出を実現できることが分かった. また, イベントの発生から 35 時間以内に約 100%のイベントを検出できることが分かった.

今後は, より大規模なデータに提案手法を適用することで, STEP1 における追跡対象の削減が, 検出効率の向上にどの程度貢献できるのかを調査する予定である. また, 本研究では使用しなかった機械学習モデルの利用や, 新たな特徴量の作成を試み, 検出精度の向上を図りたい.

謝　　辞

実験データを提供していただいた Castillo 氏に深謝する. 本研究は JSPS 科研費 19K12230 の助成を受けたものである.

文　　献

[1] P. Domm. False Rumor of Explosion at White House Causes Stocks to Briefly Plunge; AP Confirms Its Twitter Feed Was Hacked, April 2013.
[2] 東京海上日動リスクコンサルティング. 災害時のデマと流言 〜ソーシャルメディア発達の背景の下で〜, June 2011.
[3] ITmedia. 豪雨被災地で「レスキュー隊装った窃盗団出現」などデマ拡散　広島県警が注意喚起. https://www.itmedia.co.jp/business/articles/1807/09/news118.html, July 2018.
[4] K. Srijan, and N. Shah. False Information on Web and Social Media: A survey. arXiv preprint arXiv:1804.08559, 2018.
[5] X. Zhou, and R. Zafarani. Fake News: A Survey of Research, Detection Methods, and Opportunities. arXiv preprint arXiv:1812.00315, 2018.
[6] A. E. Lillie, and E. R. Middelboe. Fake News Detection using Stance Classification: A Survey arXiv preprint arXiv:1907.00181, 2019.
[7] C. Castillo, M. Mendoza, and B. Poblete. Information Credibility on Twitter. In WWW, pages 675-684, 2011.
[8] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying Misinformation in Microblogs. In EMNLP, pages 1589-1599, 2011.
[9] F. Yang, Y. Liu, X. Yu, and M. Yang. Automatic Detection

of Rumor on Sina Weibo. In the ACM SIGKDD Workshop on Mining Data Semantics, 2012.

[10] J. Ma, W. Gao, Z. Wei, Y. Lu, and K. Wong. Detect Rumors Using Time Series of Social Context Information on Microblogging Websites. In CIKM, pages 1751-1754, 2015.

[11] M. Mathioudakis and N. Koudas. TwitterMonitor: Trend Detection over the Twitter Stream. In SIGMOD, pages 1155-1158, 2010.

[12] C.J. Hutto, and E. Gilbert. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In ICWSM, 2014.