点過程を用いたフェイクニュース検出

村山　太一† 　若宮　翔子† 　荒牧　英治†

† 奈良先端科学技術大学院大学
〒 630—0192 奈良県生駒市高山町 8916-5
E-mail: †{murayama.taichi.mk1,wakamiya,aramaki}@is.naist.jp

あらまし　SNS の普及とともに,「フェイクニュース」と呼ばれる誤った情報が多く投稿・共有されており，社会問題となっている.「フェイクニュース」に対抗するために，SNS 上の投稿からフェイクニュースかどうかを判定する研究が多く行われている．本研究では，これまで着目されていなかった SNS 上での投稿数の時間的変遷に着目し，フェイクニュースを判定する手法を提案する．フェイクニュース関連の投稿数の時系列では複数回のバースト現象が生じており，この現象をフェイクニュースの検出に取り入れることは有用であると考えられる．本研究では，この時系列を点過程を用いて人々のニュースに対する関心度合いを表す “infectiousness values” に変換した時間的特徴量と，既存の投稿内容に基づく特徴量と投稿者に基づく特徴量の 3 つの特徴量を組み合わせたフェイクニュース検出モデルを提案する．

キーワード　フェイクニュース，点過程，Hawkes 過程，機械学習，時系列処理

1 はじめに

Twitter や Facebook といった Social Networking Service (SNS) は，コミュニケーションや情報交流の場として積極的に用いられており，多種多様なニュースが溢れている．Per Research Center の調査によると，アメリカの 20 歳以上の 68%が SNS からニュースを取得している [1]．SNS から容易に多くのニュースを受け取れるようになった，その一方で，ニュースの中には「フェイクニュース」と呼ばれる誤った情報もあり，多く共有されてしまっている．“フェイクニュース” という用語は多くの人々に用いられているが，明確な定義は存在しない．狭い定義として「意図的に広められた検証可能な誤った記事」が用いられている [3]．この定義は意図的であることが強調されているが，文脈から意図があるかどうかを判断することは難しい．また，フェイクニュースの検出問題において意図に関係なく誤った情報を検出することは重要である [2]．フェイクニュースの広い定義としては「ニュースの背後に存在する意図などに関わらずメディアなどを通して拡散される記事などのこと」が用いられている [4]．本論文では，後者の定義を用いる．

フェイクニュースの拡散は，SNS のプラットフォームだけでなく，民主主義を脅かす深刻な社会問題の一つとなっている．例えば，2016 年の米大統領選では，真実とは異なる各候補者にとって有利となるニュースが SNS 上で 3700 万回以上共有され，選挙結果に大きな影響を与えたといわれている [5]．

これらの問題に対して，Snopes.com [1]や PolitiFact.com [2]といったサイトが拡散された噂の検証を行っている．これらの検証サイトは手動で判断していくことからも，多くの人手や時間が必要という欠点も存在する．一方で，SNS の投稿から機械学習でフェイクニュースを検知するといった研究も多く行われている．テキスト情報 [6] やユーザ情報 [7]，拡散ツリー [9] やフォロー/フォロワーのネットワーク [10] など様々な情報が検知のために用いられている．

上記の特徴以外にも，投稿数の時系列変化もフェイクニュースの検知において有用であると考える．これは，[11] で Bots はフェイクニュースの初期の拡散に影響を与え，通常のニュースの拡散とは異なる動きが見られると報告されているためである．それにも関わらず，ほとんどの研究では「時間とともにその情報に対する注目はどのように変化するか？」といった点について考慮されずにフェイクニュースの検出を行っている．

本論文では，時間経過とともに変化するニュースへの注目度合い，つまり，時間的特徴をフェイクニュース検出のモデルに組み込む．注目度は,「ツイートが投稿された時間（投稿時間）」と「そのツイートを見る可能性があるユーザ数（フォロワー数）」を入力として点過程によって算出される．このニュースの注目度合いは，新たなユーザが情報を共有する確率に基づいて測定できることから，本研究では “infectiousness values” と呼ぶ．Infectiousness values はそのニュースに対する関心の度合いとして捉えることができ，通常この値は SNS で拡散されるニュースの特性から，時間とともに減少していくことが観察される．しかし，フェイクニュースの infectiousness values は通常の変動とは異なり，バーストが 2 度以上存在すると考える．最初のバーストは，通常のニュースの特性として SNS 上にニュースが現れることで引き起こされ，それ以降のバーストは，ニュースを知ったユーザが虚偽であることを疑ったり修正したりする投稿によって引き起こされると考えられる．

Infectiousness values の強みの 1 つとして頑健性が挙げられる．フェイクニュース検出に用いられるテキスト情報やユーザ情報などの既存特徴の多くは，拡散する人に大きく依存する．例えば，テキスト情報などは拡散したい人によって簡単に

1 : https://www.snopes.com/
2 : https://www.politifact.com/

操作され，情報やユーザ同士の関係はプラットフォームの規制やアカウントの停止などによって常に変化していく．一方で，infectiousness values は初めの投稿の動きだけでなく，時間ごとのニュース拡散の動きから算出され，一部のユーザによって操作するのは難しい．拡散ツリーも同様に人為的な操作に耐性があるが，これを取得することは infectiousness values と比較して高コストである．

本論文では，SNS 上のフェイクニュースを検出するために，既存の特徴であるテキスト情報とユーザ情報を Attention により組み合わせ，さらに時間的特徴として infectiousness values を加えたモデルを提案する．予備調査として，時間的特徴から通常のニュースとフェイクニュースを区別可能かどうか調査する．実験では，時間的特徴がフェイクニュースの検出に役立つかを実証する．

本論文の主な貢献は，

（1） フェイクニュースとフェイクでないニュース（以降，実際のニュースと呼称）の間の infectiousness values の違いを確認するという新しい問題に取り組み，フェイクニュース検出モデルにおける有効性を検討した点

（2） テキストとユーザを効果的に組み合わせ，infectiousness values を利用したフェイクニュース検出のためのマルチモデルを提案した点

（3） 3 つのデータセットで実験を行い，フェイクニュース検出における提案モデルの有効性を示した点

である．

2 関 連 研 究

本研究は SNS 上におけるフェイクニュースの検出を行う．初期の多くの研究ではテキストから抽出された言語的特徴に基づいてフェイクニュースの検知を行っている．例えば，[13] は，特殊文字，感情判定や絵文字など様々な言語機能を用いて検出を行った．[14] は，URL やハッシュタグを特徴とし，Support Vector Machine (SVM) でモデルを作成した．最近の研究では，言語的特徴を捉えるために Recurrent Neural Networks (RNN) [6] や Convolutional Neural Networks (CNN) [15] などが用いられている．[17] では，主張部分を捉えるエンコーダーとコメント情報を捉えるエンコーダーの 2 つから構成される検出モデルを作成し，主張がどの程度信頼できるかを求めた．[16] は，ニュース情報の構造を捉えるモデルを作成し，フェイクかどうかの判定を行った．

ユーザ情報を用いたフェイクニュース検知の例として，[7], [13] はフォロー／フォロワー数，登録年数などの特徴に基づいて検知モデルを作成した．最近では，ニュース記事とユーザの関係から，記事の信頼性を測定する研究が行われている [10].

SNS の投稿に対する返信や共有の伝搬ツリーを用いた検出方法を採用している研究も存在する．[18] は，疫学モデルを用いて情報カスケードを特徴量に変換しモデルを構築した．[19] は，各投稿の伝搬ツリー構造の類似度を測定するグラフカーネルに基づいた SVM で分類器を構築した．

言語やユーザ情報のみではなく複数の特徴を組み合わせてフェイクニュースを検出するマルチモデルの手法も存在する．例えば，[22] はテキスト情報とユーザの行動の組み合わせ，[21] は投稿のテキストと画像の組み合わせにより，フェイクニュースの検出を行っている．

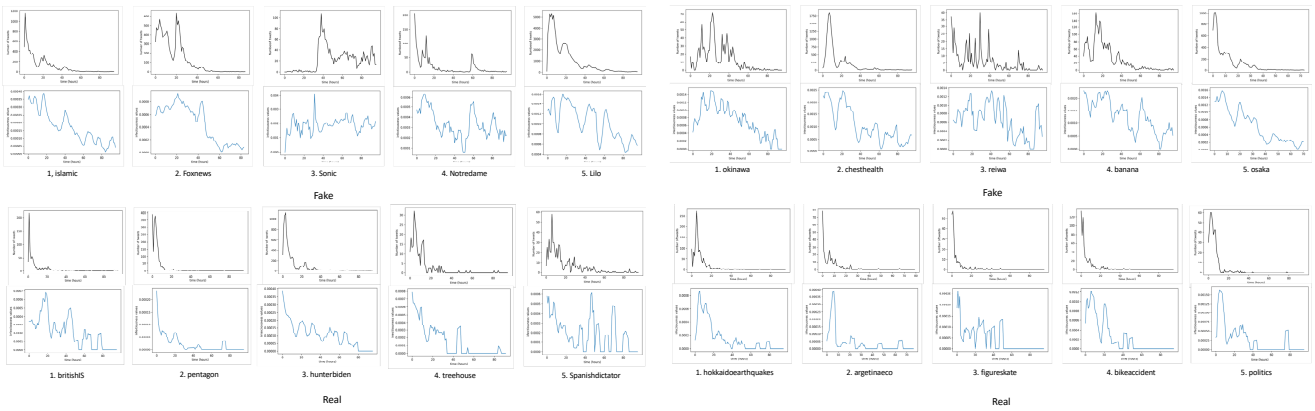
我々のモデルは，テキストとユーザ情報の関係を効果的に捉えるため，Contextual Inter-model Attention (CIM) [23] を用い，更に点過程を用いて時間的特徴を取得し，これらを用いてフェイクニュースかどうかの検出を行う．本研究に近いものとして，[12] はテキスト，ユーザ，ネットワークなどの特徴に加えて，長期的に流行する噂を検出するため，SpikeM [20] を用いて時系列の挙動を獲得し，これらを特徴量として噂かどうかの判定を行っている．本研究では時間的な特徴がフェイクニュースの早期発見にも有用であることを実証し，テキスト・ユーザ・時間の 3 つの特徴を用いてフェイクニュースの検知を行う新たなマルチモデルを提案する．

3 予 備 調 査

本章では，ニュースがフェイクかどうかを判定するタスクにおいて，投稿数の変遷などの時間情報が有用であるかどうかを検証する．最近のフェイクニュースを取得するため，日本とアメリカでそれぞれ 2018 年から 2019 年に投稿されたフェイクニュースとフェイクでない実際のニュースを 5 件づつ選択し，Twitter API [3]を用いて各ニュースに関連する投稿を取得する．アメリカのフェイクニュースの収集には，Snopes.com と PolitiFact.com で取り上げられた記事に関するキーワードや URL を用いる．日本には主要な検証サイトが存在しないため，メディアや公的機関などが否定した出来事から抽出した URL やキーワードを用いて収集を行う．フェイクでない実際のニュースに関しては，各言語の主要メディアのニュース記事の URL を用いて関連する投稿を収集する．

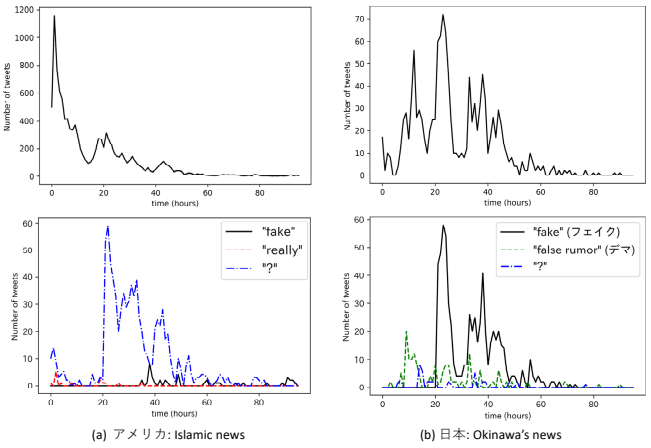
図 1 にアメリカと日本におけるフェイクとフェイクでないニュースに関する時系列を示す．各ニュースには 2 つの時系列があり，上のグラフは 1 時間ごとの投稿数の時系列を示し，下のグラフは点過程によって計算された infectiousness values の時系列を示す（計算手法については 4 章を参照）．フェイクでないニュースに関する投稿数の時系列では，最初の数時間に大きなバーストがあり，時間とともに急速に減衰していく現象が見られる．一方，フェイクニュースに関する投稿数の時系列では，最初の数時間に大きなバーストがあり，約 1 日後に再度投稿数が急増する様子が多く見られる．このことが，後者の infectiousness values の時系列に不安定な振動が生じさせる．以前の研究 [12] では，噂と非噂の投稿に対する長期の観察（56 日間）によって，噂の時系列は複数のバースト現象が観察された．一方で，本結果は短期的な観察（4 日間）でも，通常のニュースと比較して，フェイクニュースの時系列では複数のバースト現象が見られることを示している．

3 : https://developer.twitter.com



(a) アメリカ　　(b) 日本

図 1: SNS 上でのアメリカ (a) と日本 (b) におけるフェイクニュース（上 2 グラフ）と実際のニュース（下 2 グラフ）に関する投稿数の時系列と投稿数に基づく infectiousness values の時系列．各ニュースにつきそれぞれ 2 つのグラフを示す．x 軸は 96 時間の観察時間，y 軸は 1 時間ごとの値を示す．各ニュースの上部のグラフは 1 時間ごとの投稿数，下部のグラフは点過程を用いて計算された infectiousness values を示す．上の図において，ほとんどの実際のニュースの投稿数は初めに 1 回バーストするのみだが，フェイクニュースの投稿数の時系列には 2 回以上のバースト現象が見られる．同様に，infectiousness values の時系列も似たような現象が観察される．



(a) アメリカ: Islamic news　　(b) 日本: Okinawa's news

図 2: フェイクニュースに関する投稿数の時系列 (a) アメリカ: Islamic news (b) 日本: Okinawa's news. 上図は図 1 と同様に，96 時間を観察期間とし，投稿数の変化を示す．下図はフェイクであることを指摘したり疑ったりした場合に用いられる単語が含まれる投稿数の変化を示した図である．(アメリカのデータでは “fake”，“really”,“?”，日本のデータでは「フェイク」,「デマ」,「?」を検索する単語として用いる) Islamic news の結果では “fake” といった単語が初めのバースト時には見られなかったにも関わらず，最初の投稿から 20 時間経過してから急増する現象が見られる．Okinawa's news に関しても「フェイク」や「デマ」といった単語が最初の投稿から 22 時間ほど経過してから急増している現象が見られる．

フェイクニュースの投稿数の時系列において複数回急増する現象において，我々は「2 度目のバーストはそのニュースに対して否定や疑問を示した投稿によって再度注目が生じたため」という仮説を持つ．この仮説を検証するために，2 度目のバーストと疑いや否定の際に用いられる用語が投稿の中に現れる時期が一致するか確認する．その結果を図 2 に示す．疑問や否定の際に用いられる単語が最初のバーストではなく，その後のバーストとともに現れるという現象が見られる．これらの結果は我々の仮説を支持するとともに，以前の研究の結果 [24] とも一致する．疑問や否定に関する投稿（図 2）によってニュースの持つ意味が変化することから，図 1 で見られたフェイクニュースの複数のバースト現象も，ニュースの意味が変化し再び注目されることで生じたと推測できる．フェイクニュースの時系列が通常のニュースのものと比較して異なるという結果は，時間的特徴がフェイクニュース検出において有用である可能性を示唆している．

4　提案モデル

3 章で時間的情報の有用性について述べたが，事前の実験によって時間的情報のみを用いてのフェイクニュース検出では十分な精度を達成できないことがわかった．よって，時間的情報に加えてテキストとユーザ情報を用いて，SNS からフェイクニュースを検出するためのマルチモデルを提案する．モデルの全体像を図 3 に示す．

4.1　問題設定

本モデルにより取り組むのは，SNS 上の特定のニュースに関する一連の投稿（ストーリー）からそこで言及されているニュースが real か fake かを判定する問題である．

あるニュース i に関する一連のストーリーを $A_i = \{a_1, a_2, ..., a_{N_i}\}$ とする．t 番目の投稿 a_t はテキスト情報 $\mathbf{l}_t$ とユーザ情報 $\mathbf{u}_t$ から構成される $(a_t = (\mathbf{l}_t, \mathbf{u}_t))$．ストーリー i の時間的情報は $\mathbf{s}_i$ と表される．ストーリー A_i には正解値としてカテゴリ変数 $\{0,1\}^\tau$ を持つラベル $L(A_i)$ が割り当てられる．本研究の目的は予測精度が高くなるような関数 $f : f(A_i, \mathbf{s}_i) \rightarrow L(A_i)$ を構築することである．

4.2　モデル構造

提案モデルは複数のモジュールから成り立つ．テキスト，ユーザ，時間モジュールでは入力情報を潜在表現に変換する．Contextual Inter-model Attention (CIM) ではテキストモジュー

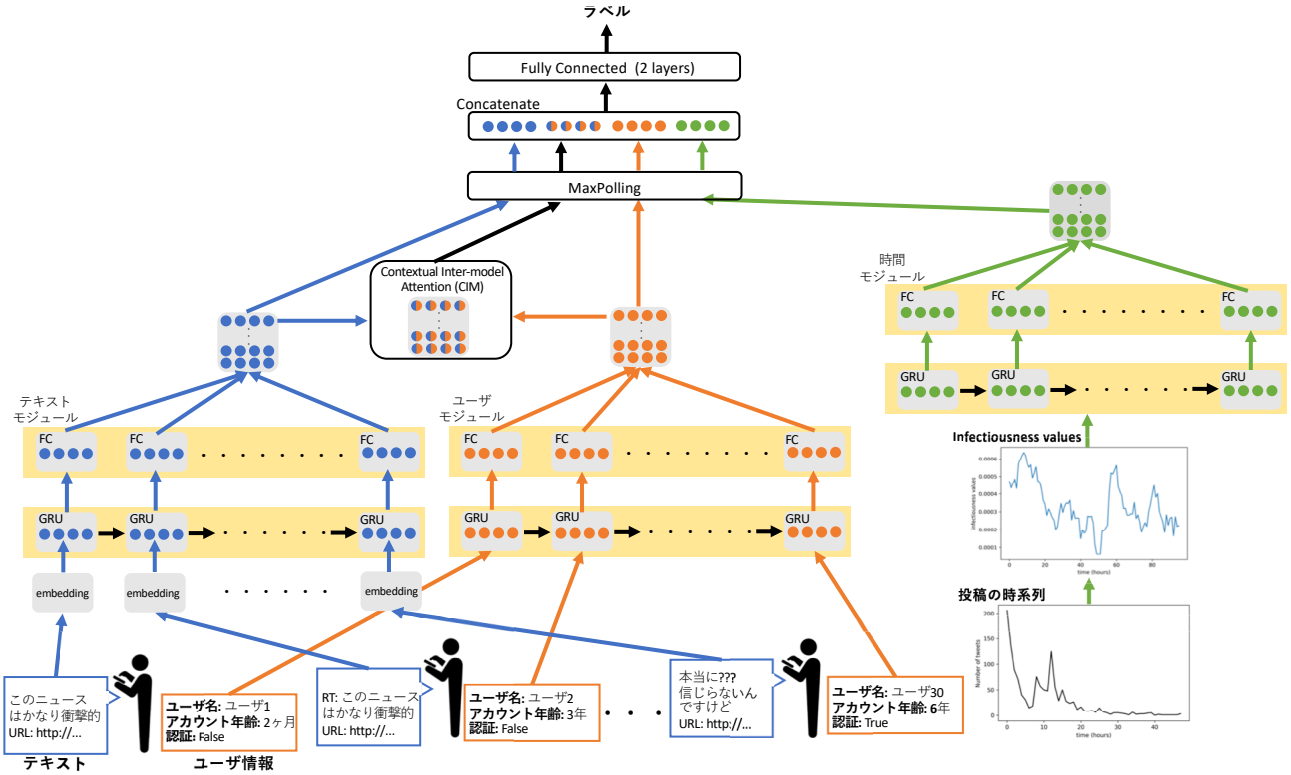


図 3: フェイクニュース検出モデルの全体像．GRU を用いて時系列順にテキスト，ユーザ情報，時間的情報の潜在表現を捉え，CIM を用いてテキストとユーザ情報を効果的に組み合わせる．出力のために各モジュール部分を連結させ，分類結果の出力を行う．

表 1: 主な記号

記号	定義
A_i	i 番目のストーリー（一連の投稿）
a_t	ストーリー A_i 内の t 番目の投稿
$\mathbf{l}_t$	t 番目の投稿のテキスト
$\mathbf{u}_t$	t 番目の投稿のユーザ情報
$\mathbf{s}_i$	i 番目のストーリーの時間的情報
$\mathbf{s}^h$	各時点の infectiousness values
$\mathbf{l}_e^t$	t 番目の投稿の I^l 次元テキスト Embedding
$\tilde{\mathbf{h}}_t^*$	各モジュールの GRU での t 番目の投稿の潜在表現
$\mathbf{h}_t^*$	各モジュールの FC での t 番目の投稿の潜在表現
$\mathbf{h}^{max\_*}$	MAXPooling を適用した各潜在表現
$\mathbf{z}$	クラス確率を表した出力
H^*	$[\mathbf{h}_t^*]$ が組み合わされた各モジュールでの出力
T^*	各モジュールの長さ
E^*	各モジュールの潜在表現 $\mathbf{h}_t^*$ の次元数
E^{con}	$\mathbf{f}^{\mathrm{l}}$ の次元数

ルとユーザモジュールで生成された潜在表現を Attention を用いて効率よく組み合わせる．最後に分類モジュールで各モジュールの出力を結合させて，分類する．各モジュールで用いる記号の一覧を表 1 に示す．

4.2.1 テキストモジュール

まず，各投稿 a_t のテキストをモデルが解釈できるように tf-idf を用いて特徴量に変換する．頻度が上位 K の単語を tf-idf に変換し，K 次元のテキスト特徴量 $\mathbf{l}_t \in \mathbb{R}^K$ を作成する．高次元でスパースなテキスト特徴量 $\mathbf{l}_t$ を低次元の潜在表現 $\mathbf{l}_t^e$ に変換するため Embedding の学習を以下のように行う．

$$\mathbf{l}_t^e = Embedding(\mathbf{l}_t), \tag{1}$$

$\mathbf{l}_t^e \in \mathbb{R}^{I^l}$ は $\mathbf{l}_t$ の Embedding を示す．

各投稿の Embedding $L_i^e = \left[\mathbf{l}_1^e, \mathbf{l}_2^e, ..., \mathbf{l}_{T^l}^e\right]$ から，テキストの潜在表現を捉えるために Gated Recurrent Units (GRUs) [25] を用いる．GRUs は Recurrent Neural Network (RNN) の一種であり長期の依存関係を捉えるために用いられ，$\mathbf{l}_t^e$ と $\tilde{\mathbf{h}}_{t-1}$ を入力として，次の潜在表現 $\tilde{\mathbf{h}}_t$ を出力とする．本モデルでは SNS 上の初期の投稿から時間の経過を考慮したテキスト特徴を捉えるために用いる．式を以下に示す．

$$\begin{aligned}
\mathbf{z}_t^l &= \sigma\left(U_z^l \mathbf{l}_t^e + W_z^l \tilde{\mathbf{h}}_{t-1}^l\right), \\
\mathbf{r}_t^l &= \sigma\left(U_r^l \mathbf{l}_t^e + W_r^l \tilde{\mathbf{h}}_{t-1}^l\right), \\
\mathbf{f}_t^l &= tanh\left(U_h^l \mathbf{l}_t^e + \tilde{\mathbf{h}}_{t-1}^l \odot W_h^l \mathbf{r}_t\right), \\
\tilde{\mathbf{h}}_t^l &= \left(1 - \mathbf{z}_t^l\right) \odot \tilde{\mathbf{h}}_{t-1}^l + \mathbf{z}_t^l \odot \mathbf{f}_t^l,
\end{aligned} \tag{2}$$

$\mathbf{z}_t^l, \mathbf{r}_t^l$ は時間 t における reset ゲートと update ゲートをそれぞれ示し，$U_z^l, U_r^l, U_h^l \in \mathbb{R}^{I^l \times E^l}$, $W_z^l, W_r^l, W_h^l \in \mathbb{R}^{E^l \times E^l}$ はそれぞれのゲートのパラメーター，E^l は GRU の出力次元数を示す．式 (2) をまとめて今後以下のように記す．

$$\tilde{\mathbf{h}}_t^l = GRU\left(\mathbf{l}_t^e\right), \quad t \in \left\{1, ..., T^l\right\}. \tag{3}$$

GRU の潜在表現 $\tilde{\mathbf{h}}_t^l$ は Fully-connected layer (FC) を適用し，

表 2: ユーザプロフィールから取得する特徴量

特徴	Type
プロフィールの長さ	Integer
ユーザの名前の長さ	Integer
フォロワー数	Integer
フォロー数	Integer
投稿数	Integer
登録期間	Integer
認証	binary
位置情報	binary

$\mathbf{h}_t^l \in \mathbb{R}^{E^l}$ が出力となる．

$$\mathbf{h}_t^l = FC(\tilde{\mathbf{h}}_t^l) \tag{4}$$

4.2.2 ユーザモジュール

ユーザ情報としてプロフィールから [7] と同様の 8 つの特徴量（表 2）を取得する．

投稿 a_t のユーザ情報である 8 つの特徴量を $\mathbf{u}_t \in \mathbb{R}^{I^u}$ と表す．テキスト特徴と同様に，GRU と FC を用いて情報の長期依存を捉えた潜在表現を求める．

$$\begin{aligned} \tilde{\mathbf{h}}_t^u &= GRU(\mathbf{u}_t), \quad t \in \{1, ..., T^u\} \\ \mathbf{h}_t^u &= FC(\tilde{\mathbf{h}}_t^u) \end{aligned} \tag{5}$$

4.2.3 時間モジュール

3 章で実際のニュースとフェイクニュースの時系列の違いについて述べた．この挙動の潜在的表現を捉えるために，点過程を元に作成された SEISMIC [26] を用いて投稿数の時系列をそのニュースが共有される確率を表した infectiousness values に変換する．SEISMIC は SNS 上のニュースの投稿数をモデル化するために，Hawkes Process [27] に基づき作成されたモデルで，本論文では時間 t までの投稿数 R_t を用いて時間 t の infectiousness values p_t とそれから求まる強度 λ_t を下記の通り算出する．

$$\lambda_t = p_t \sum_{t_i \leqq t, i \geqq 0} n_i \phi\left(t - t_i\right), \quad t \geqq t_0. \tag{6}$$

$$\phi\left(s\right) = \begin{cases} c & if \quad 0 < s \leqq s_0, \\ c(s/s_0)^{-(1+\theta)} & if \quad s > s_0, \end{cases} \tag{7}$$

n_i はニュースを閲覧する可能性のある人数（SNS 上ではフォロワー数），$\phi(\cdot)$ は情報を受け取ってから共有するまでの遅延を定量化したメモリーカーネルを示す．これらのパラメーターは [26] に則り，s_0 を 5 分，θ を 0.242，$c = 6.27 \times 10^{-4}$ とする．λ_t はこれまでの観測点に依存することから，このような形は *self-exciting* と呼ばれることが知られている．

時間とともに変化する p_t の推定は，最近の投稿に対して大きな重みを付け，古い投稿に対して小さな重み付けを行うカーネル $K_t\left(s\right)$ に依存し，以下の式で求められる．

$$p_t = \frac{\sum_{i=1}^{R_t} K_t\left(t - t_i\right)}{\sum_{i=0}^{R_t} n_i \int_{t_i}^{t} K_t\left(t - s\right) \phi\left(s - t_i\right) ds} \tag{8}$$

$$K_t\left(s\right) = \max\left\{1 - \frac{2s}{t}, 0\right\}, \quad s > 0. \tag{9}$$

infectiousness values p_t は各投稿の時間とフォロワー数を用いて求められる．本論文では，$\mathbf{s}_i$ は $\{..., \left(time_t, follower_t\right), ...\}, t \in \{1, ..., N\}$ といった $time_t$ で構成される最初の投稿から順に投稿時間とフォロワー数を示したリストとして定義される．本モデルの入力として，$\mathbf{s}_i$ を用いて各時点の infectiousness values $\mathbf{s}^h = \left(\mathbf{s}_1^h, \mathbf{s}_2^h, ..., \mathbf{s}_{T^s}^h\right)$ に変換し，テキスト情報やユーザ情報と同様に，GRU と FC を用いて時間的情報の潜在表現を求める．

$$\begin{aligned} \mathbf{s}^h &= Convert\left(\mathbf{s}_i\right) \\ \tilde{\mathbf{h}}_t^s &= GRU(\mathbf{s}_t^h), \quad t \in \{1, ..., T^s\} \\ \mathbf{h}_t^s &= FC(\tilde{\mathbf{h}}_t^s) \end{aligned} \tag{10}$$

4.2.4 Contextual Inter-model Attention

テキストとユーザの 2 つの要素で 1 つの投稿は成り立っており，これらには依存関係が存在する．しかし，GRU のみではこれらの依存関係を捉えることができない．これらの依存関係を捉えるために pair-wise contextual inter-model attention mechanism (CIM) [23] を用いる．

フェイクニュース検知において，各投稿のコンテキスト情報を活用するために，テキスト特徴量 $H^l = [\mathbf{h}_1^l, \mathbf{h}_2^l, ..., \mathbf{h}_{T^l}^l] \in \mathbb{R}^{T^l \times E^l}$ とユーザ特徴量 $H^u = [\mathbf{h}_1^u, \mathbf{h}_2^u, ..., \mathbf{h}_{T^u}^u] \in \mathbb{R}^{T^u \times E^u}$ の出力間のアテンションを算出する．なお，E^l と E^u，T^l と T^u は同じ長さである．まず，マッチング行列のペア $M_1, M_2 \in \mathbb{R}^{T^l \times T^u}$ を以下のように求める．

$$M_1 = H^l \cdot {H^u}^{\mathsf{T}}; \quad M_2 = H^u \cdot {H^l}^{\mathsf{T}} \tag{11}$$

次に，ソフトマックス関数を用いて投稿のコンテキストにおける Attention の重みを捉えるため，それぞれのマッチング行列 M_1 と M_2 から確率分布スコア $N_1, N_2 \in \mathbb{R}^{T^l \times T^u}$ を算出する．これらを Modality-wise attentive 表現と呼ぶ．

$$\begin{aligned} N_1\left(i,j\right) &= \frac{e^{M_1(i,j)}}{\sum_{k=1}^{T^l} e^{M_1(i,k)}}, \quad fori, j = 1, ..., T^l \\ N_2\left(i,j\right) &= \frac{e^{M_2(i,j)}}{\sum_{k=1}^{T^l} e^{M_2(i,k)}}, \quad fori, j = 1, ..., T^l \\ O_1 &= N_1 \cdot H^u, \quad N_2 \cdot H^l \end{aligned} \tag{12}$$

最後に，それぞれの要素から Attention を求めるため，アダマール積を算出し，H^l と H^u の Attention 表現を獲得するためにこれらを連結する．

$$\begin{aligned} A_1 &= O_1 \odot H^l, \quad A_2 = O_2 \odot H^u \\ H^{ul} &= concat\left[A_1, A_2\right] \in \mathbb{R}^{T^l \times 2E^l} \end{aligned} \tag{13}$$

4.2.5 分類モジュール

各モジュールから潜在特徴量を獲得した後に，重要な特徴を捉えるために MaxPooling を適用し，これらを連結した 1 つのベクトル $\mathbf{f}^{\mathbf{l}} \in \mathbb{R}^{E^l + E^u + 2E^l + E^s}$ を求める．

表 3: データセット

	Weibo	Twitter15	Twitter16
全ニュース数	4664	1479	813
ラベルが “true” の数	2351	371	204
ラベルが “fake” の数	2313	363	205
ラベルが “unverified news” の数	-	373	205
ラベルが “debunking” の数	-	372	199
訓練セットの数	2973	942	517
検証セットの数	525	167	97
テストセットの数	1166	370	204

$$
\begin{aligned}
\mathbf{h}^{max\_l} &= MaxPooling\left(H^l\right) \\
\mathbf{h}^{max\_u} &= MaxPooling\left(H^u\right) \\
\mathbf{h}^{max\_s} &= MaxPooling\left(H^s\right) \\
\mathbf{h}^{max\_ul} &= MaxPooling\left(H^{ul}\right) \\
\mathbf{f}^1 &= concat[\mathbf{h}^{max\_l}, \mathbf{h}^{max\_u}, \mathbf{h}^{max\_ul}, \mathbf{h}^{max\_s}]
\end{aligned} \tag{14}
$$

それぞれの特徴量の関係を捉えるために，活性化関数として $ReLU$ を採用した 2 層の Fully-connected layers (FC) を通す．最後に，各ストーリーの τ クラス分の確率を求めるため，$Softmax$ 関数を通して最終出力 $\mathbf{z} \in \mathbb{R}^\tau$ を求める．

$$
\begin{aligned}
\mathbf{f}^2 &= ReLU\left(FC\left(\mathbf{f}^1\right)\right) \\
\mathbf{z} &= Softmax\left(FC\left(\mathbf{f}^2\right)\right)
\end{aligned} \tag{15}
$$

5 実　　験

5.1 データセット

Weibo データセット [6]，Twitter15 と Twitter16 [8] の 3 つのデータセットを用いてフェイクニュース検出の実験を行う．フェイクニュースに関する各データセットは Weibo [4]や Twitter [5] といった主要な SNS から収集された投稿によって構成される．Weibo データは “true” か “fake” の 2 つのラベルが，Twitter データセットには “true”，“fake”，“unverified”，“debunking of fake” の 4 つのラベルがアノテーションされている．実験のために，各データセットを訓練，テスト，検証セットに分割する．具体的には，各データセットを 3:1 の割合で訓練とテストセットに分割し，訓練データの 15%を検証セットとして設定する．表 3 に各データセットの詳細を記す．Twitter データセットは，投稿固有の ID が公開されておりそれを元に投稿を収集するものとなっている．そのため，公開設定の変更や削除によって取得できなかった投稿もあり，既存研究のデータセットよりもサイズが少なくなっている．

5.2 比 較 手 法

提案モデルの比較手法として，フェイクニュース検出問題のために作成されたモデルを用いる．比較モデルを以下に示す．

- **SVM-TS** [28]: SNS 上のコンテキスト情報．例えばテキストやユーザ情報や，平均リツイート数などの拡散に関する特徴量などを入力として線形 SVM で構築されたモデル．
- **CSI** [22]: CSI はテキストや返信，ユーザの行動といった特徴量を用いて構築された深層学習モデル．ユーザの行動から各ニュースの傾向を算出し，各ストーリーの真偽の判定を行うモデル．
- **GRU-2** [6]: GRU-2 は各投稿のテキスト情報を入力とし，投稿のシーケンス情報を捉えるために 2 層の GRU を用いて判定を行うモデル．
- **PPC** [7]: PPC はユーザ情報を特徴量として用い，RNN と CNN で潜在表現を獲得し，判定を行うモデル．
- **提案モデル (w/o CIM)**: 提案モデルの CIM の有効性を検証するために，提案モデルから CIM を削除したモデル．
- **提案モデル (w/o time)**: 時間的情報の有効性を検証するために提案モデルから時間モジュールを除き，テキストとユーザの 2 つの特徴から判定を行うモデル．
- **提案モデル (freq)**: Infectiousness values の有効性を検証するために，各時間の投稿数を infectiousness values に変換せず，投稿数をそのまま特徴として用いたモデル．

5.3 実 験 設 定

提案モデルによる訓練セットでの予測結果が，2 値/カテゴリ損失関数を最小化するように学習する．訓練時はモデルのパラメーターは勾配ベースの Adadelta [29] を適用して更新する．さらに，過適合を避けるためにドロップアウト [30] を $\tilde{\mathbf{h}}_t^*, \mathbf{h}_t^*, \mathbf{f}^1, \mathbf{f}^2$ に 0.5 として適用する．エポック回数は 500 回として，過学習を防ぐために early stopping を 10 エポックごとに検証セットの損失を用いて適用する．

ネットワークの構造とハイパーパラメーターの設定は以前の研究 [6], [7] に基づいて設定される．まず，テキスト情報の tf-idf 変換のために用いる用語数 K を 5000 に，tf-idf から変換される Embedding の次元数 I^l を 100 に設定する．ユーザ情報の Embedding の次元数は，表 2 に基づき 8 に設定する．テキストとの特徴量に適用される GRU の長さ T^l，T^u は，既存研究 [7] に基づき Weibo データセットでは 30，Twitter15/16 データセットでは 40 として学習を行う．3 章の予備調査で，ほとんどのフェイクニュースの時系列が 1 回目のツイートが投稿されてから約 1 日経過して 2 回目の急増がみられた．このことから，時間的特徴量に適用される GRU の長さ T^s は最初の 2 日間を対象とし，1 時間ごとの infectiousness values を入力とするため 47 と設定する．各 GRU の出力サイズ (E^l, E^u, E^s) は 16，32，64，128 から FC の出力サイズ $\mathbf{f}^2$ は $E^{con}, \frac{E^{con}}{2}, \frac{E^{con}}{4}, \frac{E^{con}}{8}$ から検証セットを用いて選択する．ただし，$\mathbf{f}^1$ のサイズである E^{con} は $E^l + E^u + 2E^l + E^s$ である．

これらのモデルの性能を検証するために Accuracy と F1-measure を用いて評価を行う．Accuracy は本問題のような分類タスクにおいては一般的に用いられている評価指標である．F1-measure もまた，不均衡データにも適用可能な評価指標で

4 : https://www.weibo.com
5 : https://twitter.com

あり再現率と適合率の調和平均によって算出される．

6 結果と考察

表 4 に結果を示す．結果から提案モデルが多くのベースラインより精度が高く，マルチモデルの手法と時間的特徴の有用性を示している．ベースラインの 1 つで，作成された多くの特徴量に基づく **SVM-TS** は，テキスト・ユーザ・時間情報など様々な特徴量を組み合わせることから比較的良いモデルとなっている．一方で，**CSI** は低い精度となっている．このモデルは訓練データから各ユーザの信頼度スコアを算出し，訓練とテストの両方に現れたユーザのスコアを利用して，フェイクかどうかを検証している．しかし，実際に用いられたデータで訓練とテストの両方に現れたユーザが少なかったことから，精度が低くなったと推定される．**提案モデル**，**GRU-2**，**PPC** といったほとんどの深層学習モデルは **SVM-TS** のような多くの特徴量を作成する機械学習モデルよりも高い精度を達成している．このことからも，フェイクニュースかどうかを判定する問題において SNS の情報から良い潜在表現を獲得するために深層学習モデルを用いるのは有用である．テキスト情報を用いた **GRU-2** とユーザ情報を用いた **PPC** は，Accuracy と F1-measure の両方で高い精度を達成した．

CIM モジュールを除いた**提案モデル (w/o CIM)** と比較して，提案モデルは “unverified” ラベルを除いた全てのデータセットで高い Accuracy と F1-measure を達成した．この結果はテキストとユーザ情報を分けて潜在表現を学習するのは十分でなく，投稿を構成しているテキストとユーザ情報の相互依存関係を考慮することがフェイクニュース検出において有用であることがわかる．時間的情報を除いた**提案モデル (w/o time)** との比較においても，提案モデルは Twitter15 の “unverified” ラベルを除いて高い精度を達成している．以前の研究 [12] では噂の検知において時系列は長期の観察期間（56 日間）において有用であったが，本実験結果では時間的特徴がフェイクニュースの検知において短い観察期間（2 日間）でも有用であることを示した．Infectiousness values を変換前の投稿頻度に置換した**提案モデル (freq)** の精度は時間的情報を除いた提案モデル (w/o time) よりも，Weibo と Twitter16 データセットにおいて高かったが，提案モデルほどの精度の向上は見られなかった．この結果は infectiousness values への変換が時間的特徴から潜在情報を獲得するのに有用であることを示している．

提案モデルは “unverified” ラベル以外の指標で最も高い精度を達成している．Accuracy において Weibo データセットでは 0.937，Twitter15 データセットでは 0.831，Twitter16 データセットでは 0.819 と最も高く，F1 スコアにおいても “true”，“fake”，“debunking” のラベルで最も高いスコアとなっている．一方で “unvertified” といった曖昧なラベルについては時間的特徴を加えても判断が難しかったと考えられる．

どれくらいの期間の時間的特徴を用いることが，フェイクニュース検出に効果的であるかを評価する． くらいではどうでしょうか？有用度というのがわかりにくいです

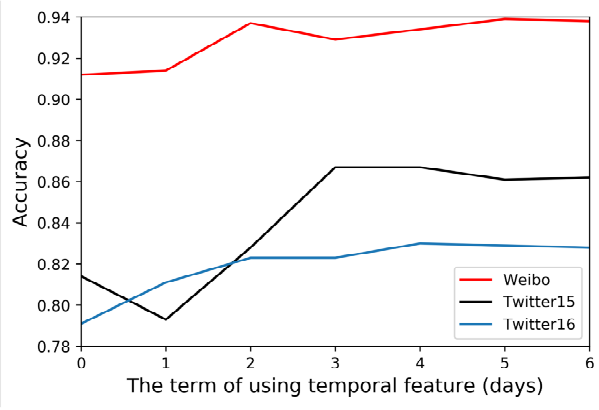


図 4: 時間的特徴の期間を変化させたモデルによる各データセットにおける精度．x 軸は 0 日 (w/o time) から 6 日間までの期間を示し，y 軸は精度を示す．

次に，どの程度の期間の時間的特徴を用いることがフェイクニュース検出に有効であるかを検証する．提案モデルに用いる時間的情報の入力期間を 0 日（提案モデル (w/o time)）から 6 日間まで変化させ，精度の比較を行う．結果を図 **??**に示す．時間的特徴を用いる期間を長くすることで精度が高くなっていくが，3 日間以上の時間的特徴を用いてもそこまで大きく精度向上しないことがわかる．今回のモデルの比較実験では 2 日間の時間的特徴を用いたが，4 日もしくは 5 日間の時間的特徴を用いることで，フェイクニュースの判定で各モデルにおいて最も高い精度を達成できることが示された．

7 おわりに

本論文のフェイクニュース検出問題に対する貢献は以下の点である．(1) 短期間の観測においても実際のニュースとフェイクニュースでは投稿数の時系列に違いが見られることを示した点．(2) テキストとユーザ情報に加え，時間的特徴として infectiousness values を組み合わせた新しいマルチモデルを提案した点．(3) 提案モデルの有効性を実験を通して示した点．しかし，時間的特徴が “unvertified” のラベルにおける有用性を示すことができなかった．今後，時間的特徴が曖昧なラベルに対しても柔軟に利用できる方法を検討する必要がある．

謝　　辞

本研究の一部は，JSPS 科研費 19K20279, 19H04221， および厚生労働省科学研究費補助金（課題番号： H30-新興行政-指定-004）の支援を受けたものです．

文　　献

[1] Shearer, E and Matsa, K, “News use across social media platforms 2018”, Pew Research Center, Journalism and Media, 2018.

[2] Takayasu, Misako, et al., “Rumor diffusion and convergence during the 3.11 earthquake: a Twitter case study.”, PLoS one, 10.4, e0121443, 2015.

[3] Shu, Kai, et al., “Fake news detection on social media: A data mining perspective.”, ACM SIGKDD Explorations Newsletter, 19.1, pp.22–36, 2017.

表 4: 各モデルによるフェイクニュース検出の結果

Method.	Weibo			Twitter15					Twitter16				
	Acc.	T F_1	F F_1	Acc.	T F_1	F F_1	U F_1	D F_1	Acc.	T F_1	F F_1	U F_1	D F_1
SVM-TS	0.827	0.831	0.837	0.599	0.772	0.598	0.608	0.544	0.574	0.743	0.488	0.551	0.549
CSI	0.780	0.750	0.803	0.556	0.601	0.631	0.550	0.530	0.507	0.552	0.511	0.475	0.443
GRU-2	0.876	0.872	0.879	0.794	0.822	0.815	0.849	0.697	0.750	0.761	0.750	**0.771**	0.723
PPC	0.914	0.912	0.917	0.806	0.748	0.840	0.807	0.730	0.778	0.803	0.760	0.711	0.767
提案モデル (w/o CIM)	0.920	0.922	0.917	0.814	0.807	0.813	**0.870**	0.745	0.791	0.850	0.782	0.747	0.791
提案モデル (w/o time)	0.912	0.913	0.910	0.814	0.857	0.806	0.868	0.677	0.791	0.864	0.829	0.717	0.776
提案モデル (freq)	0.921	0.931	0.908	0.807	0.872	0.815	0.828	0.660	0.805	0.864	0.801	0.740	0.699
提案モデル	**0.937**	**0.937**	**0.936**	**0.831**	**0.880**	**0.850**	0.833	**0.758**	**0.819**	**0.870**	**0.831**	0.739	**0.841**

[4] Sharma, Karishma, et al, "Combating fake news: A survey on identification and mitigation techniques.", ACM Transactions on Intelligent Systems and Technology (TIST), 10.3, 21, 2019.

[5] Budak, Ceren, "What happened? The Spread of Fake News Publisher Content During the 2016 US Presidential Election.", The World Wide Web Conference, pp.139–150 , 2019.

[6] Ma, Jing, et al, "Detecting rumors from microblogs with recurrent neural networks.", In Proc. of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp.3818–3824, 2016.

[7] Liu, Yang, and Yi-Fang Brook Wu., "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks.", In Proc. of AAAI Conference on Artificial Intelligence., pp.354–361, 2018.

[8] Ma, Jing, Wei Gao, and Kam-Fai Wong., "Detect rumors in microblog posts using propagation structure via kernel learning.", In Proc. of the Association for Computational Linguistics, pp.708–717, 2017.

[9] Wu, Ke, Song Yang, and Kenny Q. Zhu, "False rumors detection on sina weibo by propagation structures.", In Proc. of international conference on data engineering, pp.651–662, 2015.

[10] Nguyen, Duc Minh, et al., "Fake news detection using deep markov random fields.", In Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.1391–1400, 2019.

[11] Shao, C, et al., "The spread of low-credibility content by social bots.", Nature communications, 9.1, pp.4787, 2018.

[12] Kwon, Sejeong, Meeyoung Cha, and Kyomin Jung, "Rumor detection over varying time windows.", PloS one, 12.1, e0168344, 2017.

[13] Castillo, Carlos, Marcelo Mendoza, and Barbara Poblete., "Information credibility on twitter.", In Proc. of the international conference on World wide web., pp.675–684, 2011.

[14] Derczynski, Leon, et al., "SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours.", In Proc. of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pp.69–76, 2017.

[15] Yu, Feng, et al., "A convolutional approach for misinformation identification.", In Proc. of the International Joint Conference on Artificial Intelligence., pp.3901–3907, 2017.

[16] Karimi, Hamid, and Jiliang Tang., "Learning Hierarchical Discourse-level Structure for Fake News Detection." In Proc. of the North American Chapter of the Association for Computational Linguistics", pp.3432-3442, 2019.

[17] Zhang, Qiang, et al., "Reply-Aided Detection of Misinformation via Bayesian Deep Learning.", In Proc. of the World Wide Web Conference., pp.2333–2343, 2019.

[18] Jin, Fang, et al., "Epidemiological modeling of news and rumors on twitter.", In Proc. of the 7th Workshop on Social Network Mining and Analysis. ACM, pp.8:1–8:9, 2013.

[19] Ma, Jing, Wei Gao, and Kam-Fai Wong., "Detect rumors in microblog posts using propagation structure via kernel learning.", In Proc. of the Association for Computational Linguistics., pp.708–717, 2017.

[20] Matsubara, Yasuko, et al., "Rise and fall patterns of information diffusion: model and implications.", In Proc. of the SIGKDD international conference on Knowledge discovery and data mining., pp.56–14, 2012.

[21] Wang, Yaqing, et al., "Eann: Event adversarial neural networks for multi-modal fake news detection.", In Proc. of the SIGKDD international conference on knowledge discovery and data mining, pp.849–857, 2018.

[22] Ruchansky, Natali, Sungyong Seo, and Yan Liu., "Csi: A hybrid deep model for fake news detection.", In Proc. of the ACM on Conference on Information and Knowledge Management., pp.797–806, 2017.

[23] Ghosal, Deepanway, et al., "Contextual inter-modal attention for multi-modal sentiment analysis.", In Proc. of the Conference on Empirical Methods in Natural Language Processing., pp.3454–3466, 2018.

[24] Shao, Chengcheng, et al., "Hoaxy: A platform for tracking online misinformation.", In Proc. of the 25th international conference companion on world wide web., pp.745–750, 2016.

[25] Cho, Kyunghyun, et al., "On the properties of neural machine translation: Encoder-decoder approaches.", arXiv preprint arXiv:1409.1259, 2014.

[26] Zhao, Qingyuan, et al., "Seismic: A self-exciting point process model for predicting tweet popularity.", In Proc. of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining., pp..1531–1522, 2015.

[27] Hawkes, Alan G, "Spectra of some self-exciting and mutually exciting point processes", Biometrika, 58.1, pp.83–90, 1970.

[28] Ma, Jing, et al., "Detect rumors using time series of social context information on microblogging websites.", In Proc. of the 24th ACM International on Conference on Information and Knowledge Management, pp.1751–1754, 2015.

[29] Zeiler, Matthew D "ADADELTA: an adaptive learning rate method", arXiv preprint arXiv:1212.5701, 2012.

[30] Srivastava, Nitish, et al., "Dropout: a simple way to prevent neural networks from overfitting.", The journal of machine learning research, 15.1, pp.1929–1958, 2014.