DEIM2021 I11-4

災害情報採集のための会話tweetを活用した分類手法

藤田　俊之†　　小林　亜樹††

† 工学院大学大学院工学研究科電気・電子工学専攻　〒167–8677 東京都新宿区西新宿 1–24–2
†† 工学院大学情報学部情報通信工学科　〒167–8677 東京都新宿区西新宿 1–24–2
E-mail: †cm20046@ns.kogakuin.ac.jp, ††aki@cc.kogakuin.ac.jp

あらまし　筆者らは，災害に関する情報抽出を検索語集合によって Twitter から行う際の事前知識や見逃しを問題と捉え，会話関係を用いたクラスタリング手法を提案し，一定の有効性を示唆する結果を得てきた．本稿では，従来手法で用いてきた Latent Dirichlet Allocation (LDA) と k-means によるクラスタリングに対して代表語の割合を用いたクラスタの自動判別を導入した LDA 分類器や，Naïve Bayes 分類器 などの分類器と会話関係を組み合わせた分類手法を提案し，災害例として台風 Hagibis を対象とした tweet 集合から各分類手法の精度や再現率を求め，有効性を評価した．

キーワード　ツイート，自然言語処理，機械学習，災害情報

1 はじめに

台風などの大規模災害時では，刻一刻と被災地の状況が変化する．例えば，土砂崩れや倒木などによる通行が困難となる道路区間の発生や，河川の氾濫などが考えられる．こういった周囲の被災状況の共有や，救助要請の発信などに Social networking service (SNS) の一種である Twitter が用いられる．Twitter は，140 文字以内の文字制限を持つ tweet と呼ばれるマイクロブログを投稿することができる．また，投稿された tweet に対して Reply をすることができ，簡単に話題に参加することができる．

そのため，Twitter を情報源とすることで，被災している人々が発信していたり，やりとりしている災害情報をリアルタイムに近い形で得ることができると考えられており，Twitter をはじめとするマイクロブログサービス上の投稿から災害時の情報を取得，分析しようとする研究は数多い．そのような研究の多くでは，検索語を指定することで tweet の採集を行っている [1]–[3] が，検索語が含まれていない tweet を見逃してしまう可能性がある．

このため，筆者らは，検索語が含まれていない tweet も含めて災害に言及している tweet の採集を目指している．テキスト分析手法として知られる LDA [4] と k-means アルゴリズムを組み合わせたアルゴリズムをベースとして，Reply でやりとりされている一連の tweet 群を一つの会話と見做し，1 会話 を 1 文書と見做す 1 会話 1 文書モデルと組み合わせることで，1 tweet 1 文書とする単純な文書モデルよりも高い分類精度が得られる可能性が示唆される結果が得られた [5]．しかし，LDA と k-means によるクラスタリングで得られる複数のクラスタから，災害に言及している文書群が属していると思われるクラスタの判別は，クラスタを代表するような語群の抽出後，ユーザによるものを前提にしていた．また，分類アルゴリズムとして良く知られる Naïve Bayes などの分類器と 1 会話 1 文書モデルの組み合わせは試されていない．

そこで本稿では 3.2 節で紹介される 1 会話 1 文書モデルと，3.3 節における分類器を組み合わせた分類手法を提案し， 各分類器と 1 会話 1 文書モデルとの相性の調査を行う．1 会話 1 文書モデルとは，Twitter 上でやりとりされる一連の Reply のやりとりを会話と見做し，一つの会話内では同一の話題について言及していると考え，1 会話を 1 文書とする文書モデルである．

また，k-means 法のような教師なし学習では，クラスタリング後の各クラスタのうち，どのクラスタが災害に関連するクラスタかは定かではない．そこで，k-means 法に対して各クラスタに含まれる災害を代表するような語（代表語）の割合を利用したクラスタの自動判別を導入する．災害や時間帯によって適切な代表語は異なる可能性が考えられるが，マイクロブログが示す準実時間性を反映したトレンド分析手法の活用が考えられる．本稿では，クラスタの自動判別と k-means 法によるアルゴリズムを用いて，台風 Hagibis に投稿された tweet を対象として，その評価も行う．

2 関連研究

本研究では，災害時に投稿された tweet 群から災害に言及している tweet のみの採集を試みる．そのために，Reply 関係に着目した 1 会話を 1 文書と見做す 1 会話 1 文書モデルを一般的な分類器と組み合わせることで，分類精度の向上を図る．

本研究が目的としている災害に言及している tweet の採集には，検索語を含む tweet を検索する方法が一般的である．Sasaki ら [2] は「地震」や「揺れ」を含む tweet を用いて，台風の進路や地震の震源地を Support Vector Machine (SVM) によって特定する手法を提案した．検索語自体を拡張する研究も行われており，湯沢ら [3] は，感動詞との共起関係を利用し，リアルタイムに近い形での災害に関連する検索語群の抽出を試みた．しかし，検索語による災害情報の採集では，検索語が含まれていない tweet を見逃してしまう可能性が考えられる．そのため，

本研究で提案する分類手法では，tweet 群を直接分類するようなテキスト分類器をベースとする．

マイクロブログ上での LDA を用いたトピック分析は数多く研究されている．tweet に LDA を適用する最も単純な方式は，1 tweet を 1 文書とする文書モデルの活用である．しかし，短文であるが故の問題を指摘されることもあり，1 ユーザの全 tweets を 1 文書として扱う Author-topic model [6] や，1 tweet は 1 トピックであるという仮説に基づく Twitter-LDA モデル [7] といった改良モデルが提案されている．しかし，同一ユーザの投稿する tweet には，時間の経過による話題の転換が考えられる．これに対し，本研究では一つの話題に留まると期待される会話に注目した 1 会話 1 文書モデルを利用する．

マイクロブログが示す準実時間性を反映したトレンド分析も様々な手法が試みられている．基本的には，古くは Kleinberg が burst と呼んだ [8] 語出現頻度の急上昇を捉えて検出している [9]–[13]．また，Zhao らは提案した Twitter-LDA モデルに基づき話題を示すキーワードを抽出する手法も提案している [14]．James Benhardus ら [15] は，Streaming API で収集した tweet 集合から単位時間毎に TF-IDF 法などを用いたトレンドワードの抽出を行った．このように，tweet からの話題抽出は多く行われており，一語程度の災害を代表するような語（代表語）を抽出する方法は様々な手法が存在する．そこで，本稿では k-means のような教師なしのクラスタリング手法に対し，代表語を含む割合を用いたクラスタの自動判別法の導入を組み合わせ，災害情報の採集のための分類手法に用いる．

3 提 案 手 法

3.1 概　　要

ここでは，提案手法の概要について説明する．まず，本研究の目的は投稿された多数の tweet から災害に言及している tweet のみを採集することである．無分別の tweet 集合に対しては，これは，災害に言及するクラスとそうでないクラスに分類する問題として定式化できる．tweet の大部分が文字列で構成されていることを利用すれば，テキスト分類器の採用が妥当である．

しかし，ここで一つ問題が生じる．Tweet は（日本語では）140 文字の上限制約が課されており，テキスト分量が小さいことに起因して，通常のテキスト処理を施すことだけでは良好な結果をもたらさないことがたびたび報告されている．Tweet 集合をテキスト分類器により分類するとしたとき，最も単純なのは，1 tweet を 1 文書とするモデルであるが，この場合，この問題を引き起こすこととなる．そこで筆者らは，かねてより，tweet 内容の意味的なつながりが，reply 関係によって結ばれた tweet 間に存在することに着目し，reply によって接続された一連の tweet 集合を会話と呼び，ひとまとまりとして扱うことを提案している．すなわち，1 会話 1 文書モデルである．先行研究では，1 会話 1 文書モデルを導入して text 量の少なさを補うことで，分類器による分類精度に好影響を与えることを示唆する結果を得ている．

そこで本稿では，1 会話 1 文書モデルについて紹介（第 3.2 節）した上で，これといくつかのテキスト分類器（第 3.3 節）とを組み合わせた分類手法を提案する．単独では分類器として機能させられない教師なしクラスタリング手法を組み合わせる場合に，災害に言及しているクラスタを判別する手法についても導入する．

提案する複数の分類手法の全体像を図 1 に示す．まず，Twitter サーバより，会話となる tweet を取得する部分である．Twitter 社が提供する API には，大きく分けて Streaming API と REST API の 2 種類があるが，いずれも，単純に会話に相当する一連の tweet 木を取得する機能を提供してはいない．そのため，広くサンプルを得るために Streaming API による tweet 取得をきっかけとして，会話となる tweet を取得する手順を要する．

ここで得られた会話 tweet 集合を共通の分類対象データとして，1 会話を 1 文書とするモデルに基づき文書として扱った上で，LDA〜k-means クラスタリングを行う LDA 分類器，単純に検索語を含むか否かで分類する検索語分類器，教師あり分類の代表として Naïve Bayes 分類器，Word2vec による分散表現を元に k-means クラスタリングを行う Word2vec 分類器の 4 種を提案する．Naïve Bayes 分類器は，教師データが必要のため，同様の前提条件では使えないが，性能比較のために取り上げている．各分類器毎に得られた分類結果のうち，災害に言及しているクラスに割り当てられた tweet 集合が，目的の災害情報採集結果となる．

なお，本論文では災害情報採集にあたって，どのような状況下でどのような分類器が有効であるかの分析も行う方針であるため，この 4 分類器による分類結果を独立に分析して，種々の性能評価も行う．

3.2 1 会話 1 文書モデル

一連の Reply のやりとりのうち，reply を行った tweet を reply tweet, reply をされている tweet を original tweet とする．Reply tweet と original tweet が連鎖している一連の tweet を一つの会話と見做す．また，tweet をノードとし，reply 元 → reply 先の reply 関係を有向辺として持つグラフをモデル化すると，根付き有向木として見做せる．また，このとき有向辺の一般的な向きとは異なり，会話木は時系列方向に沿う有向辺となるように向きを揃える．

また，tweet t は表 1 に示す情報を含むと定義する．Tweet t は inreplyto 属性で参照される tweet を親とする reply 関係から会話木を再帰的に構成する．会話木に属する tweet のうち，tweet t が reply 先を持たない場合や，reply 先 tweet が観測できない場合，tweet t.inreplyto 属性は null とし，会話木の根とし，root tweet とする．会話木の構造例を図 2 に示す．1 つの会話木に属する全ての tweet を 1 つの文書と見做す文書モデルを 1 会話 1 文書モデルとする．

3.3 分 類 器

3.3.1 LDA 分類器

LDA 分類器では，各文書を bag of words 表現を利用し TF-

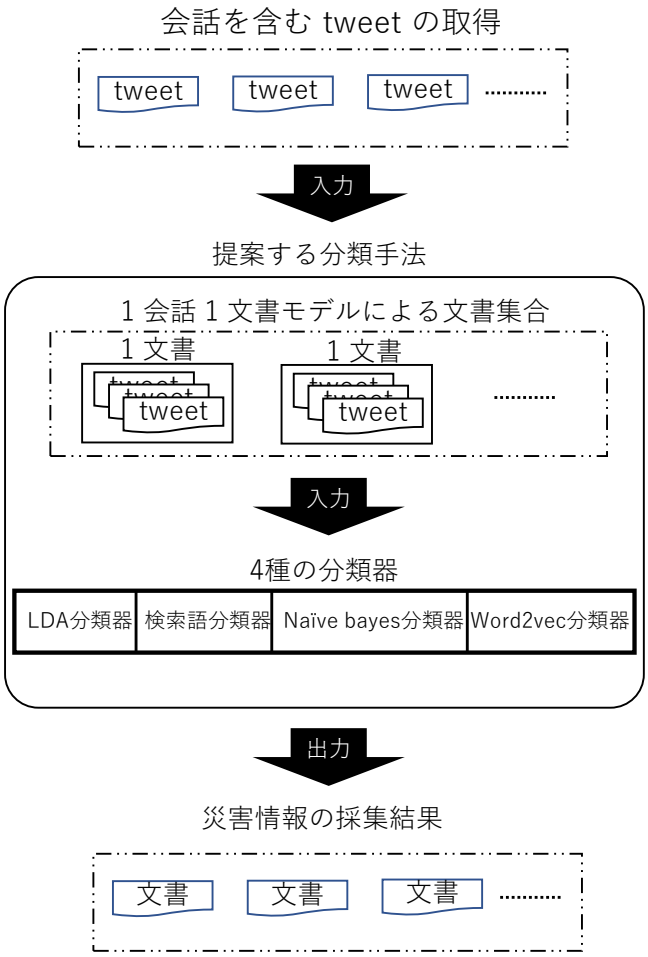

図 1　提案手法の概要

表 1　Description of the attributes of t.

Notation	Meaning
id	tweet id
timestamp	tweet の投稿時刻
text	tweet のテキスト
account	tweet を投稿したユーザの ID
inreplyto	Reply 先の tweet ID

IDF ベクトル化し，これを tweet の特徴ベクトルとする．各文書の特徴ベクトルをコーパスとして構築する LDA トピックモデルから，各文書ごとのトピック分布を得る．文書のトピック分布が点在するトピック空間を対象として k-means によるクラスタリングを行うことで，K 個のクラスタを得る．

得られた K 個のクラスタから，代表語を含む割合を用いて災害に関連するクラスタの自動判別を行う．まず，クラスタ $k(k = 1, \ldots, K)$ に属する N_k 個の文書のうち代表語が含まれる文書数を $df_{k,\text{代表語}}$ とする．各クラスタのうち，式（1）で求められる代表語を含む割合 α_k が最も高い値を持つクラスタを k' とし，$Q = \{k \mid 0.9\alpha_{k'} \leqq \alpha_k \leqq \alpha_{k'}\}$ となるクラスタ集合 Q を災害に関連するクラスタとして選択する．このクラスタ集合 Q に属する各クラスタの文書集合をまとめて災害情報の採集結果として用いる．

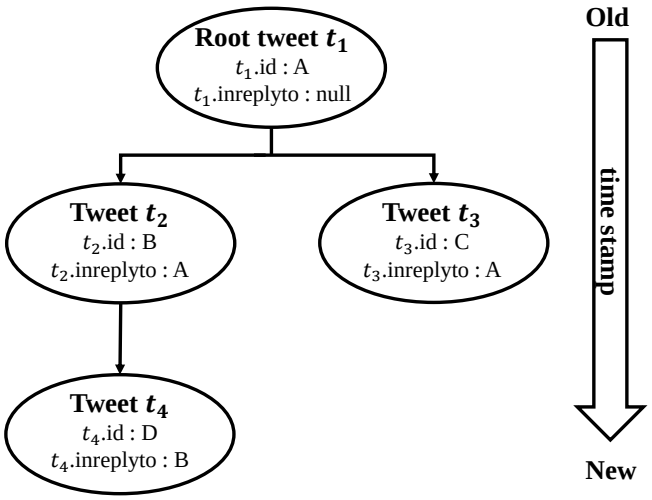

図 2　会話木の構造例

$$\alpha_k = \frac{df_{k,\text{代表語}}}{N_k} \tag{1}$$

この代表語を含む割合 α_k によるクラスタの自動判別は，災害に言及している文書のうち代表語を含む文書しか属していないクラスタが存在していた場合に見逃してしまう場合が考えられる．しかし，検索語分類器のようなものと組み合わせることで回避できると思われるため，問題は無いと思われる．

3.3.2　検索語分類器

単純に検索語を含む文書のみを検索する一般的な手法である．ここでは代表語を検索語として，検索語を含む文書のみを検索する手法を，検索語分類器と呼ぶ．多くの関連研究は，人手により多数の検索語を用意しており，本手法はそのうち，特に代表語のみである場合に相当する．検索語を含まない文書は見逃してしまうため，多くの検索漏れをもたらすものと予想される．

3.3.3　Naïve Bayes 分類器

Naïve Bayes 分類器は，ベイズの定理を使った教師あり学習を行う機械学習の一つである Gaussian naïve Bayes (GNB) を用いる分類器である．事前に，ラベル付き文書をトレーニングデータとして学習させる必要があり，本研究では，災害に言及している文書が属するクラス T と災害に言及していない文書が属するクラス F に分類されたラベル付き文書が格納されている正解 dataset（第 4.2.3 項で紹介される）を利用することで学習を行う．LDA 分類器と同様に，各文書を bag of words 表現を利用し TF-IDF ベクトル化したものを特徴ベクトルとし GNB の入力として用いる．Naïve Bayes 分類器では，GNB によって分類された文書のうち，災害に言及している文書が属するクラス T にカテゴライズされた文書を災害情報の採集結果として用いる．

3.3.4　Word2vec 分類器

Word2vec 分類器は，Word2vec [16] による分散表現と k-means を組み合わせた分類器である．文書単位での分類を行うために，式（2）のように，文書 d_i に属する j 番目の Word2vec によって得られる語ベクトル $\vec{w_j}$ の総和を，その文書の特徴ベクトル $\vec{d_i}$ とする．

$$\vec{d_i} = \sum_{j=1}^{|d_i|} \vec{w_j} \quad (2)$$

LDA 手法と同様に，この特徴ベクトルを用いて，k-means によるクラスタリング後，式（1）によって求められる代表語を含む割合 α_k を用いたクラスタの自動判別を行う．

3.4 前 処 理

本節では 3.3 節で説明した分類を実施するに当たり，共通に処理する事前処理について説明する．

3.4.1 tweet 集合の取得

会話を含む tweet 集合の取得のため，Twitter 社の提供する API のうち，Streaming API (statuses/sample)[1]と，Lookup API (statuses/lookup) [2]を用いる．

まず，Streaming API を利用し tweet 集合 T_S を取得する．tweet 集合 T_s は，投稿された全 tweet の 1%で構成される．次に，LookupAPI に対するクエリとして tweet 集合 T_s から取得した t.inreplyto を指定する．こうして得られた tweet から再帰的に reply tweet を取得し，tweet 集合 T_L を得る．こうして，会話を含む tweet 集合 $T_S \cup T_L$ の取得を行う．

3.4.2 BOT の除外

Twitter 上に投稿される tweet には，自動的に生成された tweet から Reply のやりとりを行う bot account が存在する．Bot account 同士による会話は，人同士の会話よりも長く reply のやりとりが続く会話が発生する傾向がある．ここでは，この傾向を利用した bot account 同士による会話を除外するために，bot account を検出する簡単な方法を説明する．

まず，人同士による会話よりも長い場合として，500 回以上 Reply のやりとりが続いた会話に参加していたユーザの ID を bot aacount であると見做す．こうして検出された bot account 集合を A_b とし，$T_S \cup T_L$ に含まれる bot account の tweet を T_b とする．Bot account によって投稿された tweet T_b を除いた tweet 集合を $(T_S \cup T_L) \setminus T_b$ とするとき，以後の分類における処理の対象となる tweet 集合 $T_S \cup T_L \setminus T_b$ とし，bot account による tweet を処理対象から除外する．また，bot account の検出のための tweet 群は，分類の対象とする tweet 集合の前日分を利用する．

3.4.3 不要語の除外

各分類器では，形態素解析器である MeCab で名詞と判定された語のみを利用する．また，図 3 のような Slothlib [17] で定義される stop word や，独自の不要語リスト (記号，絵文字など) に該当する語は除外する．

4 評　　価

4.1 概　　要

本章では，災害例として 2019 年に発生した台風 Hagibis に

1：https://developer.twitter.com/en/docs/tweets/sample-realtime/api-reference/get-statuses-sample

2：https://developer.twitter.com/en/docs/tweets/post-and-engage/api-reference/get-statuses-lookup

あそこ, あたり, あちら, あっち, あと, あな, あなた, あれ, いくつ, いつ, いま, いや, いろいろ, うち, おおまか, おまえ, おれ, がい, かく, かたち, かやの, から, がら, きた, くせ, ここ, こっち, こと, ごと, こちら, ごっちゃ, これ, これら, ごろ, さまざま, さらい, さん, しかた, しよう, すか, ずつ, すね, すべて, ぜんぶ, そう, そこ, そちら, そっち, そで, それ, それぞれ, それなり, たくさん, たち, たび, ため, だめ, ちゃ, ちゃん, てん, とおり, とき, どこ, どこか, ところ, どちら, どっか, どっち, どれ, なか, なかば, なに, など, なん, はじめ, はず, はるか, ひと, ひとつ, ふく, ぶり, べつ, へん, べん, ほう, ほか, まさ, まし, まとも, まま, みたい, みつ, みなさん, みんな, もと, もの, もん, やつ, よう, よそ, わけ, わたし, ハイ, 上, 中, 下, 字, 年, 月, 日, 時, 分, 秒, 週, 火, 水, 木, 金, 土, 国, 都, 道, 府, 県, 市, 区, 町, 村, 各, 第, 方, 何, 的, 度, 文, 者, 性, 体, 人, 他, 今, 部, 課, 係, 外, 類, 達, 気, 室, 口, 誰, 用, 界, 会, 首, 男, 女, 別, 話, 私, 屋, 店, 家, 場, 等, 見, 際, 観, 段, 略, 例, 系, 論, 形, 間, 地, 員, 線, 点, 書, 品, 力, 法, 感, 作, 元, 手, 数, 彼, 彼女, 子, 内, 楽, 喜, 怒, 哀, 輪, 頃, 化, 境, 俺, 奴, 高, 校, 婦, 伸, 紀, 誌, レ, 行, 列, 事, 士, 台, 集, 様, 所, 歴, 器, 名, 情, 連, 毎, 式, 簿, 回, 匹, 個, 席, 束, 歳, 目, 通, 面, 円, 玉, 枚, 前, 後, 左, 右, 次, 先, 春, 夏, 秋, 冬, 一, 二, 三, 四, 五, 六, 七, 八, 九, 十, 百, 千, 万, 億, 兆, 下記, 上記, 時間, 今回, 前回, 場合, 一つ, 年生, 自分, ヶ所, ヵ所, カ所, 箇所, ケ月, ヵ月, カ月, 箇月, 名前, 本当, 確か, 時点, 全部, 関係, 近く, 方法, 我々, 違い, 多く, 扱い, 新た, その後, 半ば, 結局, 様々, 以前, 以後, 以降, 未満, 以上, 以下, 幾つ, 毎日, 自体, 向こう, 何人, 手段, 同じ, 感じ

図 3　Stop word

投稿された tweet を対象とし，3.3 節の各分類器と 1 会話 1 文書モデルを組み合わせた分類手法の精度，再現率を用いた分類性能評価によって 1 会話 1 文書モデルとの相性の良い分類器の調査を行う．4 種の分類器のうち，Naïve Bayes 分類器は他の 3 種の分類器とは異なり教師有りの分類器であるが，分類性能の比較のため用いている．このとき，提案手法のうち，文書モデルである 1 会話 1 文書モデルの効果を分析するため，文書モデルのみ 1 tweet 1 文書とした単 tweet モデルを組み合わせた分類手法 4 種を比較手法として用いる．

また，LDA 分類器における代表語を含む割合によるクラスタ自動選択が災害に関連するクラスタだけの選択が適切に判別できているかどうかを，クラスタの自動判別を行う前の各クラスタごとに二項検定における検定量 z を用いて確認する．例えば，自動判別されたクラスタ Q のうち，統計量 z が低い災害に関連しないクラスタが混ざっていたり，自動判別されたクラスタよりも高い統計量 z が存在した場合，適切に判別されていないと判断する．また，統計量 z が高いクラスタだけが自動判別されたクラスタ Q に含まれていた場合，適切に判別されていると判断する．

以降は，各分類器，文書モデルとの組み合わせは分類器の手法名の末尾にそれぞれの文書モデルの略称を併せて表記する．例えば，LDA 分類器と 1 会話 1 文書モデルとの組み合わせを LDA 分類器 + conver, 1 tweet 1 文書との組み合わせを LDA 分類器 + single のような形で示すものとする．

4.2 実　　験

4.2.1 条　　件

実験の実行環境を表 2 に示す．Word2vec 分類器や，LDA 分類器で用いる k-means のクラスタ数は 6 とし，LDA 分類器における LDA のトピック数は 10 とした．これは，いくつかのパラメータを組み合わせした実行結果を確認し，良いと思われる組を選択した．また，Word2vec 分類器では Word2vec の教師データとして，日本語版 Wikipedia と，分類対象となる文書集合を対象とした．このとき，Skip-gram モデルと Negative Sampling モデルを利用し，表 3 に示すパラメータ群を用いた．*window* は，前後いくつの語を教師データとするか，*size* は分散表現の次元数，*negative* は Negative Sampling におけるサンプリング数である．これらのパラメータを指定し，Python の gensim ライブラリを用いて LDA や k-means, Word2vec の

実装をした．他のパラメータはデフォルトの設定としている．

表 2　実行環境

Python	3.7.6
gensim	3.8.0
MeCab	0.996
mecab-ipadic-neologd	mecab-ipadic-2.7.0-20070801-neologd-20200709

表 3　Word2vec の各パラメータ

パラメータ名	値
window	5
size	100
negative	5

4.2.2　対象 tweet

台風 Hagibis が伊豆半島に上陸したと考えられる 2019 年 10 月 12 日 19 時から同日 20 時までの期間に Streaming API で取得した tweet 集合 T_S を用いる．また，LookupAPI で取得される tweet 集合 T_L は，2019 年 10 月 12 日 18 時から同日 20 時までの期間に限定する．今回の実験では，1 会話 1 文書モデルの組み合わせによる分類精度の調査を目的としており，傾向を観察するためには 60 分程度の短期間を対象とすれば十分だと考えられるためである．

また，日本語 tweet を対象とするため，言語の属性が ‘ja’ である tweet のみを利用する．tweet には表 4 に示す 4 種の種別があると定義し，本実験では reply-tweet, quoted-retweet, normal-tweet の 3 種を用いる．

表 4　Tweet の種別

種別	説明
Reply-tweet	tweet に対して返信を行った tweet
Retweet	リツイート
Quoted-retweet	Retweet に加えて，リツイートを行った投稿者のテキストが付与された tweet
Normal-tweet	以上 3 つのどれにも属さない tweet

4.2.3　正解 dataset

4.2.1 項で定義される実験対象とする tweet 集合から，1 会話 1 文書モデルや単 tweet モデルそれぞれで構成される文書集合から 500 件づつ無作為抽出した文書集合を対象とし，各文書が 3 人の実験参加者に災害に言及しているかどうかの判定を依頼した．多数決によって各文書の正解ラベルを決定し，正解 dataset として用いる．

正解 dataset の内訳を表 5 に示す．また，多数決によって決定された正解ラベルと，各評価者で決定されたラベルの類似度を，Cohen’s kappa 係数 [18] を用いて表 6 に示す．κ 係数の解釈は表 7 に示す Landis and Kochch ら [19] によるものを踏襲しており，ほぼ全ての組み合わせにおいて，いずれの文書モデルにあっても評価者間でのラベル付け結果は概ね一致しており，ラベル付けが妥当に行われているといえる．

また，Naïve Bayes 分類器の分類精度の評価のため，正解 dataset をもとに 2 分割交差検証を用いる．このとき，正解 dataset を 2 分割して得られた各 dataset をそれぞれ datasetA, datasetB として用いる．datasetA, dataseB の内訳は表 8 のようになっており，dataseA, datasetB ともにトレーニングデータとテストデータはそれぞれ 250 個ずつ持つ．

表 5　正解 dataset 内訳

文書集合	T	F	計
1 会話 1 文書モデル	135	365	500
単 tweet モデル	73	427	500

表 6　多数決と各評価者の κ による一致度

	Labels by Majority Vote			
	文書集合 (単 tweet モデル)		文書集合 (1 会話 1 文書モデル)	
評価者	κ	解釈	κ	解釈
A	0.66	Substantial agreement	0.84	Almost perfect agreement
B	0.93	Almost perfect agreement	0.87	Almost perfect agreement
C	0.87	Almost perfect agreement	0.87	Almost perfect agreement

表 7　Cohen’s kappa κ の解釈

κ	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

4.2.4　精度と再現率

各分類手法による採集結果から，ラベル付けされた文書をもとに精度 P, 再現率 R を求める．精度 P は，文書集合 D のうち，クラス T に属する文書数 D_T とクラス F に属する文書数 D_F を用いて式（3）のように求められる．また，再現率は正解 dataset におけるクラス T の文書数 T_N を用いて式（4）のように求められる．

$$P = \frac{D_T}{D_T + D_F} \tag{3}$$

$$R = \frac{D_T}{T_N} \tag{4}$$

また，LDA 手法では，災害に言及していると判別されたクラスタが二つ以上ある場合，それらのクラスタをまとめて採集結果として用いる．また，1 会話 1 文書モデルや，単 tweet モデルで文書化した文書集合では，表 5 より，T, F それぞれに属する文書数の割合が異なるため，注意が必要である．

表 8　datasetA, datasetB の内訳

		文書集合 (1 会話 1 文書モデル)		文書集合 (単 tweet モデル)	
		トレーニングデータ	テストデータ	トレーニングデータ	テストデータ
dataset A	T	76	59	41	32
	F	174	191	209	218
dataset B	T	59	76	32	41
	F	191	174	218	209

4.2.5　二項検定の統計量 z

二項検定の z を，クラスタの分類精度の指標の一つとして用いる．各クラスタにおいて，帰無仮説 H_0 を無作為抽出であるとし，対立仮説 H_1 を無作為抽出ではないとする．また，このとき有意水準を 5%とし，棄却域を $|z| > 1.96$ とする．

統計量 z は式（5）のように定義される．ここで，X がクラス T に属する文書の数，p は正解 dataset におけるクラス T に属する文書の割合，n はクラスタ k に属するラベル付き文書の数である．この統計量 z が正の方向に高いほど災害に関連するクラスタであり，負の方向であれば災害に関連しないクラスタと判断する．

$$z = \frac{X - np}{\sqrt{np(1-p)}} \tag{5}$$

4.3 結　　果

4.3.1　クラスタ自動判別法

LDA 分類器や Word2vec 分類器における，クラスタ自動判別を行う前の各クラスタの内訳を表 9, 10 に示す．まず，LDA 分類器に注目すると，最も台風を含む割合が大きいクラスタ k' は LDA 分類器 + conver ではクラスタ $k = 2$ の $\alpha_2 = 0.01202$, LDA 分類器 + single ではクラスタ $k = 3$ の $\alpha_3 = 0.00078$ であった．クラスタ自動判別される対象となるクラスタ集合 Q に属するクラスタは，LDA 分類器 + conver では $0.010818 < \alpha_k < 0.01202$ を満たすクラスタ k であるからクラスタ $k = 2$ が選ばれ，LDA 分類器 + single では $0.000702 < \alpha_k < 0.00078$ を満たすクラスタ k であるからクラスタ $k = 3$ が災害に関連しているクラスタとして判別される．

統計量 z に注目すると，LDA 分類器 + conver, LDA 分類器 + single で判別されたクラスタの統計量 z はそれぞれ 7.86, 3.27 であるため，帰無仮説 H_0 が棄却でき無作為抽出ではないと言えるクラスタである．また，LDA 分類器 + conver, LDA 分類器 + single におけるそれぞれのクラスタ群のうち，最も高い統計量 z を持つクラスタが判別されている．

同様に，Word2vec 分類器におけるクラスタ自動判別の対象となるクラスタ集合 Q は，Word2vec 分類器 + conver では $0.739505 < \alpha_k < 0.82167$ を満たすクラスタ $k = 3$, Word2vec 分類器 + single では $0.072288 < \alpha_k < 0.08032$ を満たすクラスタ $k = 5$ である．Word2vec 分類器 + conver では最も高い統計量 z を持つクラスタが選ばれており，適切なクラスタを判別していると考えられる．また，Word2vec 分類器 + single では統計量 z が最も高いクラスタ $k = 2$ は選ばれていないが，次点で統計量 z が高いクラスタが選ばれている．

これらの結果より，単純な代表語の割合による判別であるが，おおよそ適切なクラスタを判別できていると考えられる．また，それぞれの分類手法において，代表語が含まれている文書のみではなく，代表語が含まれていない文書も併せて採集できていた．

本稿では教師無しクラスタのクラス分けに，台風を含む割合 α_k を用いた．一方，このような相対値ではなく，クラスタに含む台風を含む文書数 $df_{k,\,台風}$ を単純に用いる方法も考えられる．単純に台風を含む文書数 $df_{k,\,台風}$ が最も高いクラスタを選ぶとすると，LDA 分類器 + single では $df_{k,\,台風}$ の値が 9 であるクラスタ $k = 2$ が選ばれる．しかし，クラスタ $k = 2$ は他のクラスタと比べ多く代表語が含まれているが F に属する文書数も多く，また，統計量 z が -2.79 と低く，災害に関連していないと考えられるクラスタである．このように，代表語を多く含んでいるが災害に関連しないクラスタが選択される場合があるため，単純な台風を含む文書数 $df_{k,\,台風}$ 値よりも台風を含む割合 α_k のような相対値を用いるべきであると考えられる．

4.3.2　精度と再現率による比較

単 tweet モデル，1 会話 1 文書モデルによる分類手法の精度，再現率の結果をそれぞれ表 11，表 12 に示す．

表 11, 表 12 より，単 tweet モデルと組み合わせた LDA 分類器 + single の精度，再現率は検索語分類器 + single による精度，再現率より低いが，1 会話 1 文書モデルと組み合わせると，LDA 分類器 + conver の再現率は検索語分類器 + conver よりも再現率が高くなっている．また，検索語分類器 + single と検索語分類器 + conver を比較すると，1 会話 1 文書モデルを組み合わせることで再現率が高くなっているが精度は下がっている．対して， LDA 分類器 + single と LDA 分類器 + conver を比較すると，1 会話 1 文書モデルを組み合わせたことで精度と再現率が両方とも高くなっている．Naïve Bayes + conver は精度や再現率が Naïve Bayes 分類器 + single よりも低くなっており，1 会話 1 文書モデルを組み合わせたことで分類精度の低下が見られる．また，Word2vec 分類器では，再現率の大きな変化は見られず，1 会話 1 文書モデルとの組み合わせにより精度の向上が見られた．これらの傾向から，今回用いた分類器の中では，1 会話 1 文書モデルは，LDA 分類器に対して最も分類精度の向上に寄与することが分かった．

5　お わ り に

本稿では，分類精度の向上のために，LDA 分類器や Naïve Bayes 分類器, 検索語分類器と 1 会話 1 文書モデルを組み合わせた分類手法の提案を行った．検索語分類器や Naïve Bayes 分類器のように，1 会話 1 文書モデルと組み合わせることで分類

表 9　LDA 分類器のクラスタリング結果 (クラスタ自動判別前)

台風 Hagibis			Cluster k						合計
			1	2	3	4	5	6	-
LDA 分類器 + conver	全文書	文書数	1148	3329	2835	1142	699	1744	10897
		$df_{k,\text{台風}}$	8	40	10	3	5	11	77
		台風を含む割合 α_k	0.00697	0.01202	0.00353	0.00263	0.00715	0.00631	-
	T	全件数	9	82	13	4	8	19	135
		$df_{k,\text{台風}}$	5	33	6	1	5	6	56
	F	全件数	54	65	124	48	21	53	365
	z		-2.27	7.86	-4.62	-3.14	0.07	-0.12	-
LDA 分類器 + single	全文書	文書数	6697	39083	6445	8598	5749	6155	72727
		$df_{k,\text{台風}}$	0	9	5	4	2	2	22
		台風を含む割合 α_k	0.00000	0.00023	0.00078	0.00047	0.00035	0.00032	-
	T	全件数	7	25	15	11	7	8	73
		$df_{k,\text{台風}}$	0	8	3	4	1	2	18
	F	全件数	31	260	33	44	29	30	427
	z		0.67	-2.79	3.27	1.13	0.82	1.13	-

表 10　Word2vec 分類器のクラスタリング結果 (クラスタ自動判別前)

台風 Hagibis			Cluster k						合計
			1	2	3	4	5	6	-
Word2vec 手法 + conver	全文書	文書数	1238	2462	600	732	1395	4470	10897
		df_k, 台風	124	389	493	82	162	586	1836
		台風を含む割合 α_k	0.10016	0.15800	0.82167	0.11202	0.11613	0.13110	-
	T	全件数	7	40	17	4	17	50	135
		df_k, 台風	3	7	16	2	4	24	
	F	全件数	42	67	5	27	51	173	365
	z		-2.00	2.42	5.31	-1.77	-0.37	-1.54	-
Word2vec 手法 + single	全文書	文書数	10147	16014	6226	2319	29854	8167	72727
		$df_{\text{台風}}$	322	680	147	87	2398	198	3832
		台風を含む割合 α_k	0.03173	0.04246	0.02361	0.03752	0.08032	0.02424	-
	T	全件数	6	22	5	0	38	2	73
		df_k, 台風	1	2	0	0	15	0	18
	F	全件数	60	82	50	15	160	60	427
	z		-1.27	1.89	-1.16	-1.60	1.83	-2.54	-

表 11　1 会話 1 文書モデルによる精度，再現率の結果

台風 Hagibis	T		F	精度	再現率	採集された文書数
	台風を含む	台風を含まない				
検索語手法 + conver	56	0	21	0.73	0.41	1836
LDA 手法 + conver	33	49	65	0.56	0.61	3329
Word2vec + conver	7	33	67	0.37	0.30	2462
Naïve Bayes + conver (datasetA)	17	22	73	0.24	0.66	-
Naïve Bayes + conver (datasetB)	7	37	174	0.26	0.58	-

表 12　単 tweet モデルによる精度，再現率の結果

台風 Hagibis	T		F	精度	再現率	採集された文書数
	台風を含む	台風を含まない				
検索語手法 + single	18	0	4	0.82	0.25	3832
LDA 手法 + single	3	12	33	0.31	0.21	6445
Word2vec + single	2	20	82	0.21	0.30	16014
Naïve Bayes + single (datasetA)	8	21	121	0.28	0.91	-
Naïve Bayes + single (datasetB)	0	27	71	0.28	0.66	-

精度が低くなった分類器もあり，LDA 分類器や Word2vec 分類器のように精度，再現率の向上が見られたケースもあった．このことから，分類器ごとに 1 会話 1 文書モデルとの相性があり，本実験で用いた分類器の中では，LDA 分類器が最も相性

が良いと考えられる結果となった．

また，クラスタの自動判別手法の提案を行った．判別方法は単純な代表語の割合によるものであったが，おおよそ分類精度が高いクラスタだけの取得ができており，適切に災害情報を含むクラスタが判別できることが分かった．

今後は，Twitter-LDA モデルなどといった改良手段をベースとした他の分類器と 1 会話 1 文書モデルとの組み合わせた分類手法の分類精度も調査する．また，分類器ごとに精度や再現率の良し悪しが存在することがわかった．そのため，いくつかの分類器を組み合わせることによる分類精度の向上についても調査していきたい．また，本稿では一定期間のうちに reply 関係で結ばれた一連の tweet を会話として抽出したが，reply 間の期間が空いた場合など異なる話題に転換していくことも考えられ，一つの会話を分割するなどの工夫の検討が必要である．

文　　献

[1] 河井 孝仁, 藤代 裕之, “東日本大震災の災害情報における Twitter の利用分析, ” 広報研究 = Corporate communication studies, Vol.17, pp.118–128, 2013.

[2] Sakaki, T., Okazaki, M., and Matsuo, Y, “Earthquake ShakesTwitter Users: Real-time Event Detection by Social Sen-sors,” Proc. 19th International Conference on World WideWeb (WWW 2010), pp.851-860, 2010.

[3] 湯沢昭夫, 小林亜樹, “マイクロブログにおける感動詞との共起を利用した検索語の抽出,” 情報処理学会論文誌データベース (TOD), Vol. 12, No. 3, pp. 1–17, 2019.

[4] David M. Blei, Andrew Y. Ng, Michael I. Jordan, “Latent Dirichlet Allocation, ” Journal of Machine Learning Research 3, pp.993–1022, 2003.

[5] T. Fujita and A. Kobayashi, “Tweet Classification Using Conversational Relationships,” The 11th International Workshop on Advances in Networking and Computing (WANC’20), Nov.2020.

[6] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths, “Probabilistic author-topic models for information discovery,” Proc. of SIGKDD 2004, 2004.

[7] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li, “Compareing twitter and traditional media using topic models,” Proc. of ECIR 2011, 2011.

[8] J. Kleinberg, “Bursty and hierarchical structure in streams,” Proc. of SIGKDD2002, pp. 1–25, 2002.

[9] M. Cataldi, L. Di Caro, and C. Schifanella, “Emerging topic detection on twitter based on temporal and social terms evaluation,” Proc. of MDMKDD ’10, pp. 4:1–4:10. ACM, 2010.

[10] Hila Becker, Mor Naaman, and Luis Gravano, “Beyond trending topics: Real-world event identification on twitter,” Proc. of ICWSM 2011, 2011.

[11] Edward Benson, Aria Haghighi, and Regina Barzilay, “Event discovery in social media feeds,” Proc. of ACL 2011, pp. 389–398, 2011.

[12] 中島伸介, 張建偉, 稲垣陽一, 中本レン, “大規模なブログ記事時系列分析に基づく流行語候補の早期発見手法”, 情報処理学会論文誌データベース（TOD）, Vol.6, No.1, pp.1–15, 2013.

[13] 鳥海不二夫, 榊剛史, “バースト現象におけるトピック分析,” 情報処理学会論文誌, Vol.58, No.6, pp. 1287–1299, 2017.

[14] Wayne Xin Zhao, Jing Jiang, Jing He, Yang Song, Palakorn Achananuparp, Ee-Peng Lim, and Xaoming Li, “Topical keyphrase extraction from twitter,” Proc. of The Anuual Meeting of the Association for Computational Linguistics 2011, pp. 379–388, 2011.

[15] J. Benhardus, Jugal Kalita, “Streaming trend detection in Twitter,” Int. J. Web Based Communities, Vol. 9, No. 1, pp. 122–139, 2013.

[16] Mikolov, Tomas, et al, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781 2013.

[17] 大島裕明, 中村聡史, 田中克己, “SlothLib:Web サーチ研究のためのプログラミングライブラリ,” 日本データベース学会 Letters, 6, 1, pp.113–116, 2007.

[18] Cohen, J, “A coefficient of agreement for nominal scales,” Educational and Psychological Measurement 20, pp.37–46, 1960.

[19] Landis, J. R. and Koch, G. G., “The measurement of observer agreement for categorical data,” Biometrics. Vol.33, pp.159–174, 1977.