修士論文

ユーザによる情報理解の支援を目的とした意見抽出システム

吉原 昂司

主指導教員 篠田 陽一

北陸先端科学技術大学院大学
先端科学技術研究科
（情報科学）

令和３年３月

概要

SNS(Social Networking Service) の普及により、簡単に一般の個人がメッセージを投稿・閲覧し、情報の共有を行うことが可能となった。さらに、SNS は緊急時において安否確認などの役目を果たすようになり、人々が生活を送る上で欠かせないものとなっている。しかし、SNS を通して発信される情報の中には信頼性に欠けるものもあり、それが急速かつ広範囲に広がっていき人々の行動に影響を与え、社会問題に発展してしまうという側面もある。

Twitter 上では特定の投稿に関するユーザ同士の議論が展開される。その中には投稿の信頼性を判断する上での重要な意見も投稿されている。こういった議論の中で発生した意見を参考にすることで投稿の信頼性を判断しようとするが、Twitter 上での膨大な投稿の中で最後まで議論を追い、重要な意見を含んだ様々な投稿を把握するのは容易では無い。従って、議論から発生した意見を自動で整理できる仕組みが必要である。

本研究は、Twitter 上のユーザ同士によって展開される、あるトピックに関する議論の中で発生した意見をカテゴリ毎に分類し、トピックに疑問や興味を持ったユーザの情報に対する理解の支援を行う。　提案するシステムは、ユーザが注目している投稿に関連する議論から発生した意見を抽出、文書間の類似度を計算しカテゴリ毎に分類する。これにより、投稿の信頼性を判断する際に必要な多面的な意見をユーザに提供することが可能となり、ユーザの情報に対する理解の向上が見込める。

実験では、Twitter 上で収集した身近で起きたニュースに関する 3 つの実データを用いて、カテゴリ毎に分類を行い文章同士の類似度を計算した。実験で得られた結果に対する考察を行い、同じカテゴリに属している文章同士の類似度を上昇させる手法の検討を行った。

類似度の比較を行うため通常の手法とは別に新たに 3 つの手法を考案し、スレッド連接法ではツイートのスレッドを 1 つの文章にすることで、スレッド特有の文章の特徴が現れ類似度の上昇を期待した。除名詞法では文章に共通で存在するトピックに関連する名詞を取り除くことで、その文章特有の特徴が現れると考えた。文節分解法では文章を分解し類似度の計算を行う。2 つの長い文章でお互いに同じカテゴリに属しているにも関わらず他の文章との類似度が高く、目的としたカテゴリの分類ができなくなるといった問題の解決を考えた。

目 次

図 目 次

表 目 次

第1章　はじめに

本章では、本研究の背景と目的、本論文の構成を述べる。

1.1　背景

SNS(Social Networking Service) の普及により、簡単に一般の個人がメッセージを投稿・閲覧し、情報の共有を行うことが可能となった。さらに、SNS は緊急時において安否確認などの役目を果たすようになり、人々が生活を送る上で欠かせないものとなっている。本研究では SNS の一つである Twitter に着目した。

2011 年 3 月 11 日に発生した東日本大震災では、地震や津波の影響で電話やメールが繋がらないといった状況の中で、多くの人々が Twitter を通じて自身の安否報告や災害情報の収集を行った。しかし、Twitter を通して発信される情報の中には信頼性に欠けるものもあり、それが急速かつ広範囲に広がっていき人々の行動に影響を与え、社会問題に発展してしまうという側面もある。記憶に新しい件では Twitter 上でのコロナウイルスによる影響で生産元である中国でのティッシュペーパーやトイレットペーパーの生産量が減少し、それに伴って国内での流通が減り品薄になるので購入した方が良いといった投稿を閲覧し影響を受けた人々によるティッシュペーパーやトイレットペーパーの買い占め問題がある。この問題により、全国のスーパーやドラッグストアでのティッシュペーパーやトイレットペーパーの品薄が発生した。

Twitter 上では特定の投稿に関するユーザ同士の議論が展開される。その中には投稿の信頼性を判断する上での重要な意見も投稿されている。先述の件でもこの投稿に関してトイレットペーパーの原料であるパルプは中国から輸入されておらず、ほとんど国産であるといった投稿がされている。この投稿を参考にすれば、問題となった投稿の内容もティッシュペーパーやトイレットペーパーの生産元が中国では無く、国産であるのならば、国内での流通には関係ないのでは無いかといっ

た見方ができる。こういった議論の中で発生した意見を参考にすることで投稿の信頼性を判断しようとするが、Twitter 上での膨大な投稿の中で最後まで議論を追うのは容易では無い。従って、議論から発生した意見を自動で整理できる仕組みが必要である。

1.2 目的

本研究は、Twitter 上のユーザ同士によって拡散される根拠のない風説 (流言) の真偽を明らかにすることではなく、その情報を受け取って判断するユーザの支援を行う。Web 情報をユーザがどう受け取り判断しているかについての研究 [1] では正しい情報を正確に定義できず、最終的な真偽の判断はユーザが行うため、正確さをシステムで完全に決定することはできないと述べている。

本研究の目的は、ユーザが注目している投稿に関連する議論から発生した意見を抽出しカテゴリ毎に分類することである。これにより、Twitter 上でのユーザが注目した投稿に対する議論に関連する意見の整理を行うことができる。よって、投稿の信頼性を判断する際に必要な多面的な意見をユーザに提供することが可能となり、ユーザの情報に対する理解の向上が見込める。

1.3 本論文の構成

1 章では研究の背景、目的を述べた。2 章では関連技術と関連研究を述べる。3 章では本研究の提案について述べる。4 章では設計・実装について述べる。5 章ではシステムの評価実験を行った。6 章では本研究のまとめを述べている。

第2章　関連技術と関連研究

本章で本研究の関連技術および関連研究について述べる。関連技術として、本研究で対象としているSNSの一つであるTwitterの機能や利用方法や情報を整理してくれているまとめサイトおよび意見抽出を行うツールについて述べる。関連研究として、Twitterにおけるユーザ支援の方法や文章分類の方法について述べる。

2.1　関連技術

2.1.1　Twitter

Twitterはマイクロブログ・ミニブログと呼ばれるSNS(Social Networking Service)の一つである。ユーザはツイート(tweet)と呼ばれる上限140文字の短いテキストや画像・動画を投稿することで他のユーザに情報を発信することが可能である。リアルタイム性が高く、他のユーザとの情報交換が容易である。ここからTwitter上での基本的な機能について説明していく。Twitterの機能であるフォロー(follow)を利用することで、自分がフォローした相手のフォロワー(follower)になり、自分の興味のあるユーザのツイートを自分のタイムラインと呼ばれる画面に表示することができる。タイムラインでは自身のツイートとフォローしたユーザのツイートが表示される。ユーザのツイートは基本的に自身のフォロワーに対して発信される。

本研究ではTwitter上での情報伝播手段として4つ挙げており、その中でも(2)リプライと(4)引用リツイートを投稿に対する意見としている。

(1) ツイート
(2) リプライ (reply)
(3) リツイート (retweet)
(4) 引用ツイート

ツイート

ユーザから投稿される 140 文字以内の短い文章。投稿したユーザやそのユーザのフォロワーのタイムラインに表示される。

リプライ

ツイート内容の最初に「@ユーザ名」から始まるツイートである。リプライ元ユーザ、リプライ先ユーザ及び両方をフォローしているユーザのタイムラインにツイートが表示される。対象としたツイートに対する自身の意見や感想が書かれているケースが多い。

リツイート

リツイートは自分の興味の持ったツイートを自分のフォロワーのタイムラインに表示させ知らせることができるツイート転送機能である。Twitter 上でのツイートの拡散を行う手段となっている。

引用リツイート

引用リツイートはリツイートと違い、リツイート対象に対して自分のツイートを追加してリツイートすることができる。リプライと同じく対象としたツイートに対する自身の意見や感想が書かれているケースが多い。

2.1.2 Togetter

Togetter[2] は、Twitter のツイートを集めて公開できるウェブサービスである。ユーザ自身が自分の気になったツイートについて関連するツイートを手動で選択しまとめることができる。閲覧したユーザはそのトピックに対してのツイートが整理されているので情報に関する全体像を短時間で知ることができる。

2.1.3 意見 (評価表現) 抽出ツール

国立研究開発法人情報通信研究機構旧知識処理グループ 情報信頼性プロジェクトによって開発されたもの [3] で、形式に沿ったテキストファイルを入力として、機械学習によって意見や評判および評価がテキスト中に出現するそれぞれの文に存在するかどうかを判定する。その文に評価情報が存在する場合は良い、悪い等の評価表現やその評価の批評、肯定的か否定的かどうか、また評価をしている主体を抽出してくれる。

2.2 関連研究

川口ら [4] の研究では、あるニュースに対してのユーザの理解の支援を目的とし、ニュースに対する反応ツイートを抽出、共にユーザに提示することでユーザに多様な視点や情報を提供するシステムを提案している。反応ツイートの評価方法と 3 つの観点から評価している。1 つ目は、反応ユーザの熟知度としてユーザの過去のツイートを収集、それらをユーザの興味としてモデル化し、反応ニュース内の記事の内容と比較している。2 つ目は、反応ユーザの信頼度として反応ユーザのフォロワー数に注目しており、フォロワー数の多さを社会的な信頼に値すると考えていた。3 つ目は反応ツイートの注目度として対象ツイートへの反応ツイートが得ている「リツイート数」や「いいね」の数をどれだけ注目されたかを示す指標としていた。

藤川ら [5] の研究では、ある情報が正しいかどうかを直接判定するのではなく、情報に対する反応を「疑いの有無」と「根拠の有無に」分類してユーザに提示することで、ユーザが情報の真偽を判断するための支援を行なっていた。

Muqtar Unnisa ら [6] は、k-means と階層的クラスタリングを用いたスペクトルクラスタリングと呼ばれるクラスタリングアルゴリズムを使用して、ツイートを賛成意見または反対意見にクラスタリングする教師なし学習アルゴリズムを提案している。

Itai Himelboim ら [7] は、Twitter のトピックを分類するために、ネットワーク全体の構造を利用している。情報の流れの特徴を示す指標として確立されている 4 つのネットワークレベルの指標（密度、モジュール性、集中性、孤立したユーザーの割合）を 3 段階の分類モデルに利用した。そして、情報フローの構造として、分

割、統一、断片化、クラスター化、イン＆アウトのハブ＆スポークネットワークという 6 つの構造を提案した。その後、トピック毎に構成しやすいネットワークの構造パターンを示していた。

第3章　提案

本章はTwitter上でのユーザ支援の現状および本研究の提案について述べる。

本節では、本研究で提案するシステムの概要を述べる。提案システムの目的は、ユーザ支援を最終目的とした意見の抽出・整理である。

3.1　ユーザによる情報理解の支援の課題

Twitter上での流言による社会問題が発生しているため、情報を理解しやすくするためにユーザを支援する研究が行われている。しかし、情報を発信したユーザの信頼性や正確性による情報の取捨選択やユーザの興味モデルに合わせた情報の提供による情報の部分的な偏りが生じている。これでは、ユーザに偏った情報しか提供する事ができず、多面的な視点を持たせる事ができない。

3.2　意思決定のプロセス

本研究では、図3.1のような4段階のプロセスモデルを定義する。プロセスの流れとして、ユーザが情報を入手してからその情報に疑問や興味を示したならば、関連する情報の収集を繰り返し、最終的に情報に対する判断を行う。

情報の理解が苦手な人は、このプロセスにおいて入手した情報に対して疑問を生じることなく判断を行ってしまったり、疑問があったとしても中々自分が必要としている情報を見つけれない。

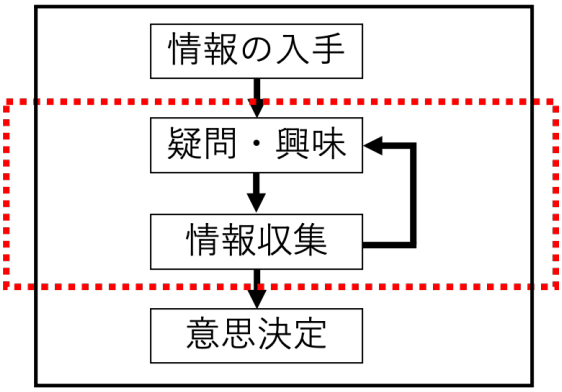

図 3.1: ユーザの意思決定のプロセス

3.3 提案：ユーザによる情報理解の支援のための意見抽出システム

本節では、本研究で提案するシステムの概要を述べる。提案システムの目的は、図 3.1 における疑問・興味の発生、情報収集の支援である。図 3.2 のようなシステムを構築することで、Twitter 上で展開されている議論の中で投稿されている様々な意見をユーザに提供する事で、ユーザに多面的な視点を持たせる効果が期待できる。自身の意見との相違により疑問を持つ事や意見のカテゴリ化による、情報収集の効率化を目指す。

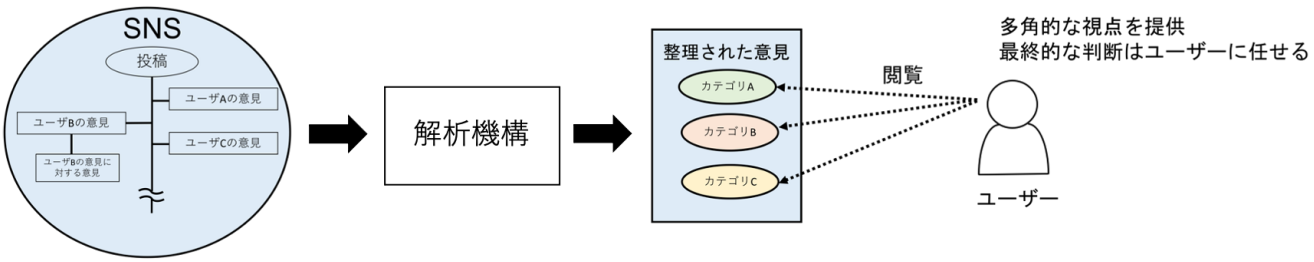

図 3.2: 提案システム

3.3.1　解析機構

手法 1:単純類似度法

単純類似度法では、図 3.3 のように、対象ツイートに対して行われた返信ツイートや引用リツイートを収集し、各ツイートの文章同士の類似度を行う。

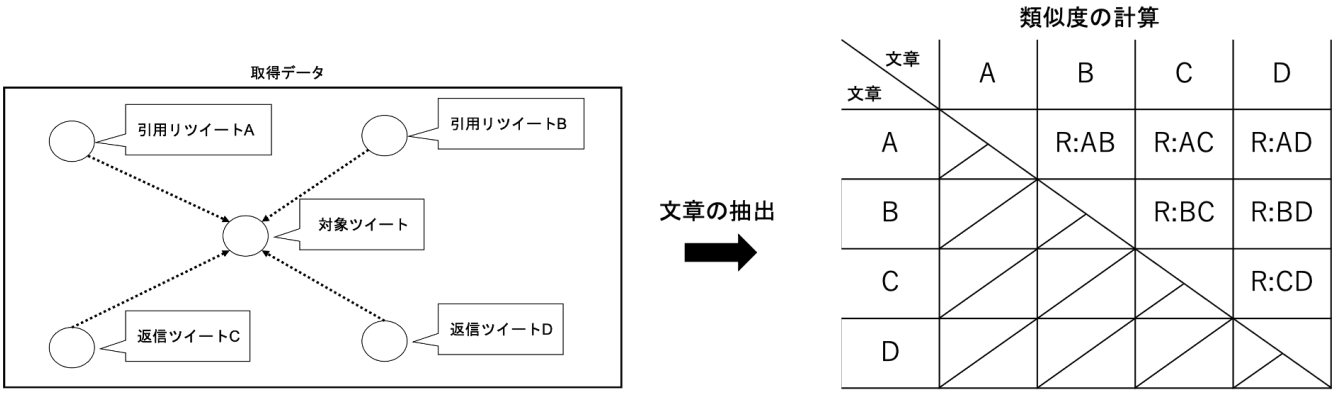

図 3.3: 手法 1：単純類似度法

手法 2：スレッド連接法

スレッド連接法では、図 3.4 のように、各ツイートのスレッドをまとめて一つの文章にし類似度の計算を行う。あるユーザが何かを疑問に思いツイートした場合に多くのケースでは、そのツイートに関する内容や答えが返信として返ってくると考えられる。よって、そのスレッド特有の文章の特徴が現れ、そのスレッドに関連する文章の類似度に影響を与えると考えた。

手法 3：除名詞法

除名詞法では、対象ツイートに出現する名詞を各ツイートから取り除いて類似度の計算を行った。これにより、各文章から対象ツイートに関連する共通した単語が無くなり、その文章特有の特徴が現れると考えた。

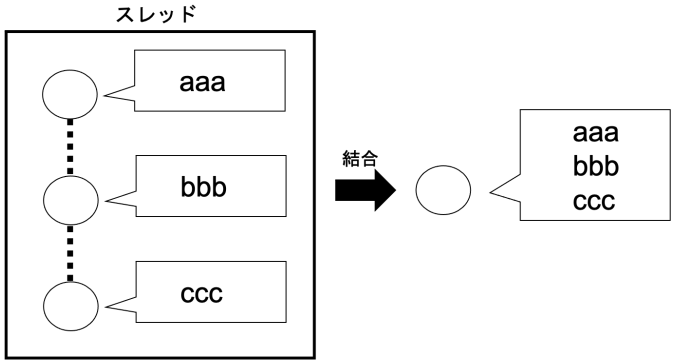

図 3.4: 手法 2：スレッド連接法

手法 4：文節分解法

文節分解法では、各ツイートの文章を規則に従って、いくつかの文章に分解し類似度の計算を行った。これにより、手法 1・2・3 では頻出頻度の低い名詞はその文章固有の特徴として計算されてしまい、他の文章と類似している部分が薄れてしまう。文章を分解し類似度を計算する事で、各文章の類似した部分を抽出する事ができると考えた。

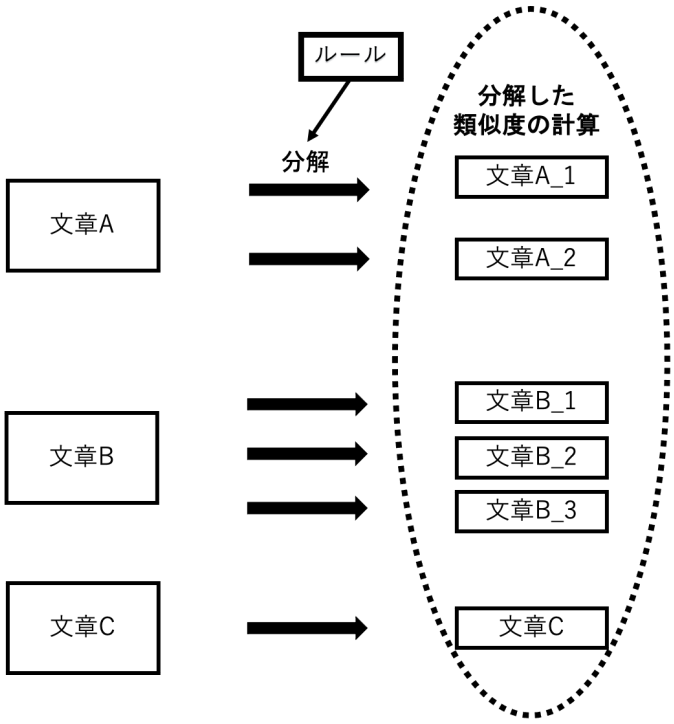

図 3.5: 手法 4：文節分解法

第4章　設計・実装

本章では、システムの設計について述べる。文書分類の手法として、ルールベースによる分類と機械学習による分類が考えられる。

4.1　文書分類の手法

4.1.1　ルールベース

ルールベースの手法では、事前にカテゴリー分けをするためのルールを設定しておく。例としては、ニュースの分類をする際に「サッカー」や「野球」などの競技名が出てきたら「スポーツ」というカテゴリーにするというような作成者の経験に基づいてルールは設定される。しかし、多種多様な文章を分類する上で新しいワードやトピックを適切に処理する上で、ルールを再設定しなければならないため手間がかかってしまう。また、作成者の意図に左右されてしまう。

4.1.2　機械学習

機械学習は学習したデータをもとに効率よく文章を分類できるのでルールベースより有効である。機械学習による文書分類には2種類の方法がある。1つが教師なし学習で文章のトピックの正解を与えずに学習を行わせ、こちらで指定したトピック数の群に分けるものである。もう一つが教師あり学習でこちらは文章のトピックの正解を与え、その決められたトピックを分類するための学習モデルを作成する。その前段階として文章のベクトル化を行う必要がある。1つがカウントベースの手法で、文書内の単語の出現頻度をもとにベクトルを算出する。もう1つが推論ベースの手法で、近い意味の単語や文などを近いベクトルに対応させる分散表現を算出するモデルを使用する。

4.2 設計

文章を分類するために、機械学習の手法をもとに自分で文章データのトピックを分類し、トピックごとの文章同士の類似度を確認するため図 4.1 のような設計とした。

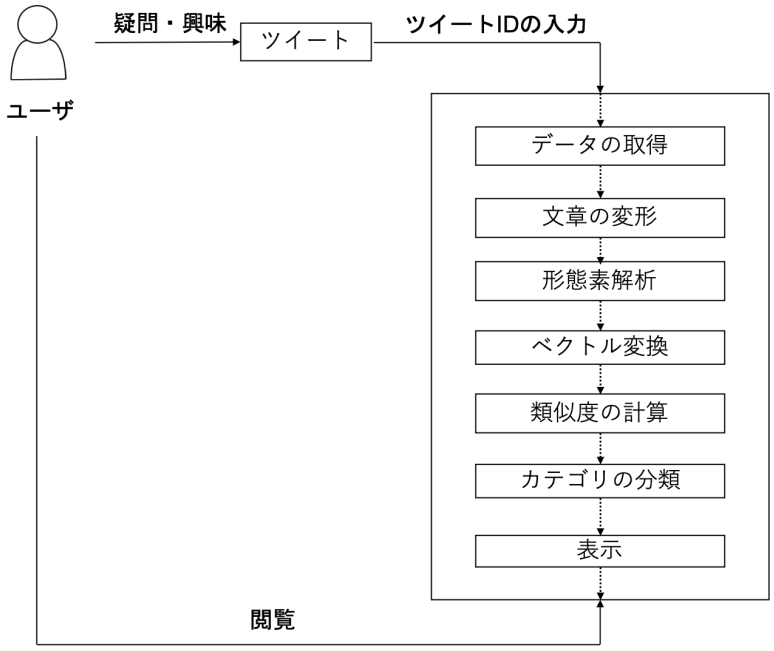

図 4.1: 意見抽出システムの設計

4.3 実装

評価実験を行うために、提案手法を実装した。提案手法の実装にはツイートのデータを取得するために Twitter API、文節分解法における文章の変形には文区切りを行えるライブラリを利用し、各ツイート文からの単語抽出には Mecab による形態素解析を用いた。そして、形態素で分離された文章を tf-idf によるベクトル変換を行い、cos 類似度で文章間のベクトルの類似度を計算した。

4.3.1 データ収集

Twitter API とは Twitter 社が提供しているサービスで、ツイートやタイムラインの取得、リツイートやいいね等の Twitter のサービスを、公式のウェブサイトを

経由せずに利用できる。この API は Twitter のアカウント情報とアプリケーションを登録することで利用できる。今回は、API の機能の内、Twitter ユーザのツイート情報を取得するのに利用した。

4.3.2 文章の変形

文章の文区切りには、ja_sentence_segmeter と呼ばれるライブラリを利用した。単純な文区切りで用いられるルールとしては、改行や「。」「！」「？」などの記号で区切られることが多いが、現実の文章ではこのような単純なルールではこちらの意図したような文に区切ることが難しい。このライブラリでは、「」や（）内に句点や感嘆符がある場合は、その文章を文の途中にある句点や感嘆符で区切ることなく処理を実行できる。

4.3.3 形態素解析

Mecab は京都大学の研究チームで開発された言語、辞書、コーパスに依存しない汎用的な設計を基本方針としたオープンソース形態素解析エンジンである。ユーザ自身が辞書やコーパスを用意することで新規語でもサポートが可能となっている。Twitter では日々、若者言葉等の流行語が頻繁に使用されている。以上の点からそれらの新規語でも認識が可能である Mecab を提案手法の実装に適した形態素解析エンジンであると考えた。Mecab の辞書にはシステム辞書とユーザ辞書があり、実装では処理が早いシステム辞書を使用している。その中でも mecab-ipadic-NEologd というシステム辞書を利用した。mecab-ipadic-NEologd は Web 上に存在する多くの言語資源から取得した新語を追加することで作成されたシステム辞書である。単語分かち書き辞書であり、IPA 辞書と呼ばれる標準的な辞書では網羅されていないネット上で流行した単語や慣用句やハッシュタグをエントリ化している。また、週 2 回以上の更新が行われているため、Twitter 上で日々使用されている流行語の形態素解析に対応できると考えた。

4.3.4　ベクトル変換

tf-idf は、Spark Jones(1972) らによって提唱され、その後 Salton and McGill らによって議論された、主に情報検索に使用される重み付き指標である。
tf は (Term Frequency) の略で、単語の文書内の出現頻度である。ある文章中に出現する頻度が多いほど、その単語は重要であり、その文章の特徴を判別するのに有用である。

$$tf = \frac{\text{文書 } A \text{ における単語 } X \text{ の出現頻度}}{\text{文書 } A \text{ における全単語の出現頻度}}$$

idf は (Inverse Document Frequency) の略で、ある単語が出て来る文書頻度の逆数となる。多くの文章で出現してくる単語は、一つの文章の特徴語にはなりづらい。逆に数少ない文書にしか出現しない単語は、その文章の特徴を判別するのに有用となる。

$$idf = \log \frac{\text{全文書数}}{\text{単語 } X \text{ を含む文書数}} + 1$$

右辺に 1 を足すことで idf が 0 にならないようにしている。

tf-idf は、この二つの概念を合わせたものである。

$$tfidf = tf * idf$$

4.3.5　類似度の計算

cos 類似度は、ベクトル空間モデルにおいて文書間の類似度を計算するのに用いられる手法である。そのまま、ベクトル同士の成す角度の近さを表現するため、1 に近いほど文書同士が類似しており、0 に近いほど類似していない事になる。

$$\begin{aligned} cos(\vec{q}, \vec{d}) &= \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| \cdot |\vec{d}|} \\ &= \frac{\vec{q}}{|\vec{q}|} \cdot \frac{\vec{d}}{|\vec{d}|} \\ &= \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \cdot \sqrt{\sum_{i=1}^{|V|} d_i^2}} \end{aligned}$$

正規化された単位ベクトルについては、以下の式で計算が可能となる。

$$cos(\vec{q}, \vec{d}) = \vec{q} \cdot \vec{d} = \sum_{i=1}^{|V|} q_i d_i$$

第5章　評価実験

本章で評価実験について述べる。

5.1　実験データ

データに記載されているツイートとそのツイートに対して行われた返信ツイート・引用リツイート、またそれらのスレッドにある非公開ツイートを除く全てのツイートを取得した。ツイート内容にあるURLの「http」・「https」が名詞として認識されURLを含む文章同士の類似度に影響を及ぼすため、URLを削除して実験を行った。その他MeCabによる形態素解析の際に「〜しないのか」という、文章中に出現する「の」という単語が名詞として定義されてしまう。今回の手法では、文章同士の共通した名詞が類似度の数値に影響するので、「の」という名詞を処理の中で抽出しなかった。

5.2　実験結果

実験1

表5.1に示した、市街地に出没した熊への対策や見解を述べているツイートに対して行われた返信ツイート・引用リツイートを収集した。そのツイートの中には、表5.2に示したような人間が熊に餌となる柿の実を与えたせいで、味を覚えた熊が人里まで降りてきたのではないかといった意見が4件存在した。文章同士の類似度を計算し、4件ある意見同士の類似度を手法毎に確認した。

- 実験データ1
  概要：市街地に出没した熊への対策と見解
  対象ツイートID：1327032121811038208

返信ツイート数：18
引用リツイート数：16
全ツイート数：34

表 5.1: 対象ツイート 1

ユーザ名	テキスト
@kumamoriTOKYO	クマ対策の基本は誘因部の除去。柿は、どうしても人家の近くにあることが多いね、、こちらの都市部でも、時々柿の木はあります、家のそばに。日本らしい風景ですね。クマも昔から日本にいる、日本の住民です。 市街地に出没のクマ 胃の内容物などの大半が柿の実 石川県

実験データ 1 の実験結果を表 5.3・5.4・5.5 に示す。

実験 2

表 5.6 に示した、北陸自動車道で発生した立ち往生についてのツイートに対して行われた返信ツイート・引用リツイートを収集した。そのツイートの中には、表 5.7 に示したような通行止めを行えばこのような被害は発生しなかったというような意見が 4 件存在した。文章同士の類似度を計算し、4 件ある意見同士の類似度を手法毎に確認した。表 5.7 に示した文章は手法 4 による文節の分解を行なったため、1 行で分解した 1 文節となっている。

- 実験データ 2
  概要：北陸自動車道に発生した立ち往生
  対象ツイート ID：1347218262078078978
  返信ツイート数：25
  引用リツイート数：111
  全ツイート数：136

実験データ 2 の実験結果を表**??**・5.8・5.9・**??**に示す。

表 5.2: 意見 1-1(人間が餌を与えたせい)

番号	ユーザ名	テキスト
3	@mimon01	果樹の果実は、果肉を狙って食べる鳥獣のためにある。果樹の方だって、なるべく省エネして、たくさんの種子を運んでほしいから、野生の果実は果肉が薄く種ばかり多い。それを品種改良して可食部分を増やしたのはヒトだから、ヒト向けの果実を野生動物に与えると、味をしめて人里に下りてくる。
5	@sunsuke3122	そりゃこんな美味しい餌あるよって教えたのち人が下山する姿見たら、その餌探して降りてくるでしょう。貧困地域でお金ばら撒いたら寄ってきたり家突き止められて強盗入られるリスク上がるのと同じ原理じゃないかと思います。
6	@mimon01	山にいる自然のクマは柿の実なんて主食にしない。それに柿の実を与えて人里に出没させたのは、たぶん熊森のせいだ。
33	@koushuuwobukkow	昔だってクマがでてこれば山狩りしてだだろうが住人なわけないだろ？仲良く農業でもしてたと思ってんのか？おまえらが栄養のある餌あげるから数年後はいまの何倍のクマが殺されるんだろうね？

表 5.3: 実験 1-1 結果

番号	番号	類似度
3	5	0
	6	0.059
	33	0
5	3	0
	6	0
	33	0.091
6	3	0.059
	5	0
	33	0.067
33	3	0
	5	0.091
	6	0.067

表 5.4: 実験 1-2 結果

番号	番号	類似度
3	5	0
	6	0.074
	33	0.017
5	3	0
	6	0
	33	0.083
6	3	0.074
	5	0
	33	0.057
33	3	0.017
	5	0.083
	6	0.057

表 5.5: 実験 1-3 結果

番号	番号	類似度
3	5	0
	6	0.068
	33	0
5	3	0
	6	0
	33	0.102
6	3	0.068
	5	0
	33	0
33	3	0
	5	0.102
	6	0

表 5.6: 対象ツイート 2

ユーザ名	テキスト
@UN_NERV	【石川県の北陸自動車道で約 90 台の車動けず】北陸自動車道の下り線で 2 台の大型車が動けなくなり、7 日夜 11 時から、石川県と富山県の間の一部の区間が通行止めになっています。現場周辺では除雪作業が進められていますが、およそ 90 台の車が動けなくなっているということです。

表 5.7: 意見 2-1(通行止めをするべきだった)

ツイート番号	ユーザ名	テキスト
13	@tsiokb	教訓から学ばないのかね？ こういうドライバーいる限り助けられない。 道路もなんで通行止めにしないのかね。 結局呼ばれる自衛隊も大変だよな
22	@kanawandy1	なんで通行止めしなかったの？ 同じ繰り返し。 死ぬよ？
52	@ho_shi_mimimi	3 年前の悪夢が再び… 先に通行止め出来なかったんかな？
55	@tatsu1000	新潟の立ち往生の報道から学べよな… 次はどこの日本海側の県で同じことやるんだか。 てか、こうなる前に通行止めにしろや。

表 5.8: 実験 2-2 結果

ツイート番号	ツイート番号	類似度
13	22	0.201
	52	0.136
	55	0.084
22	13	0.201
	52	0.274
	55	0.170
52	13	0.136
	22	0.274
	55	0.189
55	13	0.084
	22	0.170
	52	0.189

表 5.9: 実験 2-3 結果

ツイート番号	ツイート番号	類似度
13	22	0
	52	0
	55	0
22	13	0
	52	0
	55	0
52	13	0
	22	0
	55	0.079
55	13	0
	22	0
	52	0.079

実験 3

表 5.10 に示した、男性を襲った熊の殺処分についての批判を行ったツイートに対して行われた返信ツイート・引用リツイートを収集した。そのツイートの中には、表 5.11 に示したような熊の保護施設の建設を求める意見が 2 件存在した。他には表 5.12 に示してある追いかけてまで処分を行う必要があったのかという疑問の声も 4 件存在している。表 5.12 に示した文章は手法 4 による文節の分解を行なったため、1 行で分解した 1 文節となっている。

- 実験データ 3
  概要：男性を襲った熊の殺処分についての批判
  対象ツイート ID：1344589454883790848
  返信ツイート数：66
  引用リツイート数：38
  全ツイート数：104

表 5.10: 対象ツイート 3

ユーザ名	テキスト
@kumamoriTOKYO	冬眠したくて必死な思いで柿を食べに来た。駆除とあるが、山中に逃げた熊の足跡を辿り、狩猟として (猟師の獲物で私物となる) 殺しました。 17 日朝４時半前南魚沼市 除雪作業準備中の男性が熊に襲われた。熊は逃げたが、猟友会が足跡を追い、男性を襲ったとみられる熊を駆除。

表 5.11: 意見 3-1(保護施設の設置)

ユーザ名	テキスト
@HaruHaru_pts	毎回思うんだけど熊森で保護施設でも作ったら済む話じゃない？
@rushifeus	熊森さん、こういう不幸な事故が起きないよう、人里から 20km くらい離れたところに『熊保護センター (生涯飼育施設)』をつくって運営してくださいな。 力加減を考慮しない大型動物との同衾はしかねますので。 エ？シブナイデシイク ?? ホントニ？

実験データ 3 の実験結果を表 5.13・5.14・5.15・5.16 に示す。

表 5.12: 意見 3-2(疑問の声)

番号	ユーザ名	テキスト
3	@05191967	逃げる熊まで殺すのがそんなに楽しいのかね？
32	@cromer_kn	令和に狩猟って、まだ縄文人？ がいるんだ。 追いかけて殺すって、サイコパスの所業。 クマさん怖くて痛くて苦しかったね、せめて安らかに眠ってね。 What a primitive and wierd bear hunter Nigata has!! It's time to think about it .
36	@nacorotta	逃げたのを追ってまで殺す必要があるのか？
40	@mayuamu2	追う必要があるの？ 撃ちたいだけでしょ

表 5.13: 実験 3-1 結果

番号	番号	類似度
3	32	0
	36	0
	40	0
32	3	0
	36	0
	40	0
36	3	0
	32	0
	40	0.7
40	3	0
	32	0
	36	0.7

表 5.14: 実験 3-2 結果

番号	番号	類似度
3	32	0.03
	36	0.046
	40	0
32	3	0.03
	36	0.048
	40	0.051
36	3	0.046
	32	0.048
	40	0.7
40	3	0
	32	0.051
	36	0.7

表 5.15: 実験 3-3 結果

番号	番号	類似度
3	32	0
	36	0
	40	0
32	3	0
	36	0
	40	0
36	3	0
	32	0
	40	0.7
40	3	0
	32	0
	36	0.7

表 5.16: 実験 3-4 結果

番号	番号	類似度
3	32_3	0
	36	0
	40_2	0
32_3	3	0
	36	0
	40_2	0
36	3	0
	32_3	0
	40_2	0.661
40_2	3	0
	32_3	0
	36	0.661

5.3 評価

本節では、提案手法に対する評価を行う。提案手法の目的である同じカテゴリに属している文章同士の類似度の増加について手法 1 の単純類似度法を基準とし、提案したその他の手法と比較して評価する。

手法 2：スレッド連接法

実験 1・実験 3 では類似度が増加するケースがあったが、逆に減少してしまうケースも発生した。このことからスレッドを連接することで二つの文章のスレッド間で関連性の発生を示唆することができる。実験 2 では同じカテゴリに属している各文章にスレッドがなく類似度の変動値は小さかった。これらの結果から、この手法を有効に利用するためには、文章にスレッドが存在していることやスレッド間での関連性が類似度を増加させるために重要である。

手法 3：除名詞法

実験 1・実験 3 の結果から手法 1 における結果に左右されやすく、手法 1 での類似度が 0 であった場合効果を期待できない。しかし、実験 2 では類似度が共通して

減少している。実験 2 の結果から、共通したトピックに関連する名詞を持っていると考えられ減少値を利用してカテゴリの分類に応用できると考えられる。

手法 4：文節分解法

実験 2 では、文章の文節を分解したことで、文章の特徴が細分化され各文章の一部の文節と大きい類似度を示した。実験 3 では、名詞の共通点が少なく文節の分解を行なっても、類似度の増加が見られなかった。これらの結果から、この手法では文章を分解しても共通の名詞が存在しなければ効果を期待できない。しかし、手法 1 において共通した名詞による類似度を確認できれば、この手法によって高い類似度を示すことができる。

表 5.17: 意見 3-2(疑問の声)

データ番号	概要	総ツイート数
1	市街地に出没した熊への対策と見解	34
2	北陸自動車道で発生した立ち往生	136
3	男性を襲った熊の殺処分についての批判	104

第6章　おわりに

6.1　まとめ

提案システムの目的である、ユーザによる情報の理解の支援を行うためにTwitter上で展開されている議論の意見を整理するために文章の分類を目指した。同じカテゴリに属している文章同士の類似度を大きくすることができれば、文章の分類を行うことができると考えまず初めに、収集したデータの意見の分類を手動で行いカテゴリ化した。

その後、分類したカテゴリに属している文章同士の類似度を計算し確認した。同じカテゴリに属している文章同士の類似度を高くするためにいくつかの手法を提案し、手法毎による類似度の値を確認した。いずれの手法でもデータ毎に類似度の変動値に差が発生した。

6.2　考察

実験1における手法1と手法2による結果の比較から文章同士の類似度の上昇値は大きくないが、手法1では類似性がない結果から手法2では類似性が発生した。このことからスレッド間での関連性があることを示唆することができる。

手法3による操作では手法1の場合と比べ類似度が大きくなるケースを確認できた。これは、対象ツイートに含まれている名詞が省かれることにより文章のベクトル変換における単語の母数が減少することが関係すると考えられる。関連する文章では単語が削除され共通した部分が無くなってしまうが、逆に関係しない文章の類似度を大きくする効果がある。

また、類似度が小さくなるケースでは、文章同士の類似度がそこまで大きくない場合に関連する単語を含んでいると考えられ、対象ツイートに関連する内容を有している可能性がある。

手法4による文章を分割し行う類似度の計算では、2つの長い文章ではお互いに同じカテゴリに属しているにも関わらず他の文章との類似度が高く、目的としたカテゴリの分類ができなくなるといった問題を解決できる。しかし、同じカテゴリに属していても短い文章同士の類似度の増加には影響しない。

6.3 今後の展望と課題

同じカテゴリに属している文章同士の類似度を高くするためにいくつかの手法を提案したが、本研究では、文章の分類まで至っていない。そこで分類を行うため計算した類似度をもとにクラスタリングを行い、同じカテゴリに属している文章同士がクラスタになるか確認を行いたい。

また、手法4による類似度の計算では分割した文章同士の類似度は高いが、文章の分類をするためには計算された類似度を元の文章の類似度に対応させる方法が必要である。手法ごとにデータによって類似度の変動値に差が生じたことからデータの特徴から類似度の上昇に適切な手法を選択できる可能性もある。

今回の実験では、データ数も少なく同じトピックに属する文章が少なかったため、データ数を大きくし、巨大なトピックが存在している場合の類似度の計算を行いたい。

謝辞

本研究を進めるにあたり、主指導教員である篠田陽一教授には適切にご指導賜りました。深く感謝いたします。また、日々の御指導だけではなく研究者としての目標やあるべき姿を見せていただき研究や物事に対する考え方の目標になりました。副指導教員である知念賢一准教授には技術的なご支援をいただきました。深く感謝いたします。また、本研究室の宇多仁助教には研究に関する活発なご指導を賜りました。深く感謝いたします。インターンシップ指導教員である丹康雄教授には中間発表などの研究の節目に客観的な立場から的確なご助言をいただき感謝いたします。WIDE プロジェクトに所属する先生方には、専門的な立場からのご意見やご指摘をいただきました。

本研究室修了生の渡邊司揮氏には、研究計画提案書や論文の添削のご支援だけではなく、研究に関しての様々なご意見をいただきました。深く感謝いたします。本研究室の博士後期課程の三浦良介氏には様々な面から有意義なご助言や研究の考え方についての活発な議論をいただきました。本研究室の博士前期課程の馬越紘氏、門脇真之佑氏、古寺雄馬氏、本間可楠氏、油布翔平氏、岡田真一氏、梅内翼氏、片岡拓海氏、瀧島和則氏には研究に関する様々な議論や研究生活を送る上での多大なご助力をいただきました。最後に家族の皆様には学生生活および私生活をあらゆる面で支えていただき感謝いたします。修士論文の提出に至るまで皆様に多大なご支援をいただき、ありがとうございました。

参考文献

[1] 山本 祐輔, ウェブ情報の信憑性分析に関する研究, 京都大学, 2011.

[2] Togetter. https://togetter.com. (参照 2021-01-27)

[3] 意見（評価表現）抽出ツール. https://alaginrc.nict.go.jp/opinion/index.html. (参照 2021-01-27)

[4] 川口 天佑, SNS におけるニュース理解の支援を目的とするツイート推薦, 九州大学, 2017.

[5] 藤川 智英, マイクロブログ上の流言に対するユーザの態度の分類, 電子情報通信学会技術研究報告. DE, データ工学, pp.55-60, 2011

[6] M. Unnisa and S. Raziuddin, Opinion Mining on Twitter Data using, Int. J. Comput. Appl., 148 (12) (2016), pp. 975-8887(2016)

[7] Himelboim I, Smith M.A, Rainie L, Shneiderman B, and Espina C. Classifying Twitter Topic-Networks Using Social Network Analysis, Social Media+ Society (2017)