高知工科大学
KOCHI UNIVERSITY OF TECHNOLOGY

災害時におけるTwitter のつぶやきデータに対するテキストマイニング

著者	宍戸 海士
発行年	2018-03
その他のタイトル	Text mining for twitter tweet data in case of disaster
URL	http://hdl.handle.net/10173/1969

平成 29 年度

学士学位論文

災害時における Twitter のつぶやきデータに対するテキストマイニング

Text mining for twitter tweet data in case of disaster

1180336　　宍戸 海士

指導教員　　清水明宏

2018 年 2 月 28 日

高知工科大学 情報学群

要 旨

災害時における Twitter のつぶやきデータに対するテキストマイニング

宍戸 海士

震災時には被災地の情報をリアルタイムに把握する必要がある．東日本大震災や，熊本地震の際には情報入手手段の一種として Twitter が用いられた．しかし，ツイートデータを扱うにはツイートデータを解析する必要がある．既存研究ではツイートデータの解析として SVM，単純ベイズを用いてツイートの分類を行なっている．その精度は 70%であり，十分とは言えない．

本論文では RNN の一種である LSTM を用いてツイート分類を行う．LSTM を用いることで既存方式より分類に必要な特徴を損なうことなく分類できることを実証する．また，LSTM がツイート分類において最も分類精度が高いことを実証する．

キーワード　テキストマイニング，Twitter，災害，LSTM

Abstract

Text mining for twitter tweet data in case of disaster

Kaito Shishido

In the event of the disaster, it is necessary to grasp the information of the afflicted area in real time. During the Great East Japan Earthquake and the Kumamoto earthquake as a kind of information acquisition means Twitter was used. However, to handle tweet data It is necessary to analyze the tweet data. In the existing research, we As an analysis, SVM, simple Bayes is used to classify tweets. Its accuracy is 70%, which is not enough.

In this paper, we classify tweets using LSTM which is a type of RNN. By using LSTM, without losing features necessary for classification than existing methods Can be classified. Also, in LSTM classification in tweets We demonstrate that classification accuracy is highest.

key words Text mining, Twitter, disaster, LSTM

目次

目次

図目次

表目次

第 1 章

はじめに

東日本大震災や熊本地震等の震災では，被災地の状況や必要な物資等の情報をリアルタイムに把握する必要がある．情報入手手段の一つとして SNS の Twitter が存在する．東日本大震災の際には停電のため，情報が入手出来なくなってしまった人々の為に Twitter によって被災地の状況や必要な物資，津波警報，避難所のマップ，救援要請等の情報が発信された [1]．熊本地震においても，地震発生時から一週間をピークに爆発的にツイートが増えている [2]．このように Twitter を用いることで被災地の情報を入手することが可能である．しかし，Twitter から被災地の情報を得るには，ツイートデータを解析する必要がある．東日本大震災の際にはツイートデータの解析を手作業で行なっていたため，必要な情報を得るまでに時間と手間がかかってしまうと言う問題が指摘されていた．この問題を解決するために既存研究では，ツイートデータの解析を行なっている [3]．ツイートデータの解析として機械学習である Support Vector Machine(以降 SVM とする) 及び単純ベイズを用いたツイートの分類を行なっている．二つの手法の分類精度は SVM が 63.3%，単純ベイズが 70%となっている．しかし，ツイートデータは膨大なため，分類精度が 70%では十分であるとは言えない．そこで，分類精度が低い理由として，既存の方式では分類するための特徴を捉えきれていないと考えた．この仮説を証明するために，本稿では自動的に特徴を求めることが可能である Neural Network(以降 NN とする) を用いてツイートの分類を行う．NN の中でも自然言語処理の分野で使用されている Reccurent Neural Network(以降 RNN とする)，その一種である Long short-term memory(以降 LSTM とする) を用いる．LSTM を用いてツイートの分類を行い単純ベイズ，SVM との比較を行い LSTM では分類の特徴を捉えることができ分類精度が向上することを実証する．また，その結果に至った原因について考察する．

第 2 章

既存研究

2.1　既存方式で用いる技術

2.1.1　SVM

SVM とは機械学習の一手法であり，主に分類や回帰といった問題解決に用いられる．SVM は「線形しきい素子」を用いて，パターン識別器を作成する手法で，与えられた学習データから「マージン最大化」の基準で「線形しきい素子」のパラメータを学習させていく手法である [4]．データは図 2.1 の様に様々な境界線を引くことが可能である．そこで，各データのもっとも近いデータとの距離を最大になる様に境界線を決定する．すると図 2.2 のようになり，最良の結果が得られる．

2.1　既存方式で用いる技術

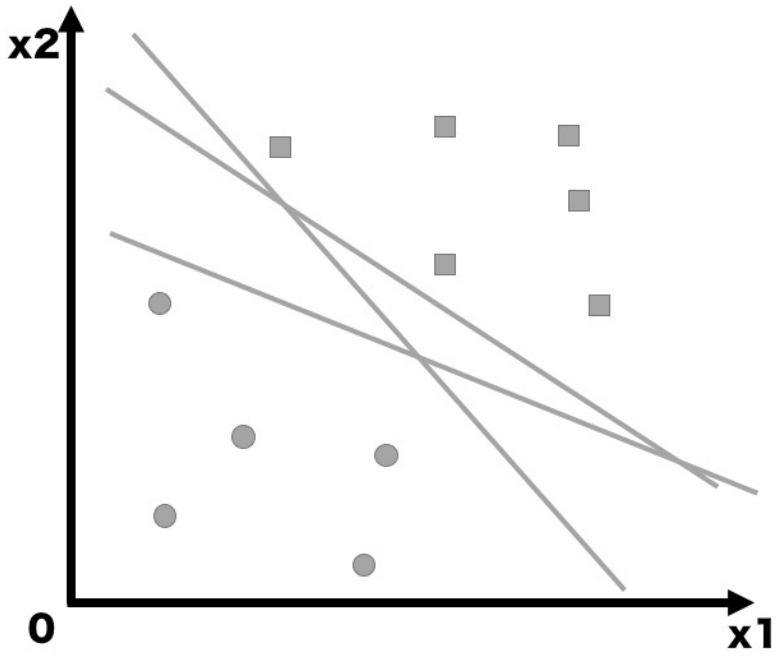

図 2.1　データの境界線 [4]

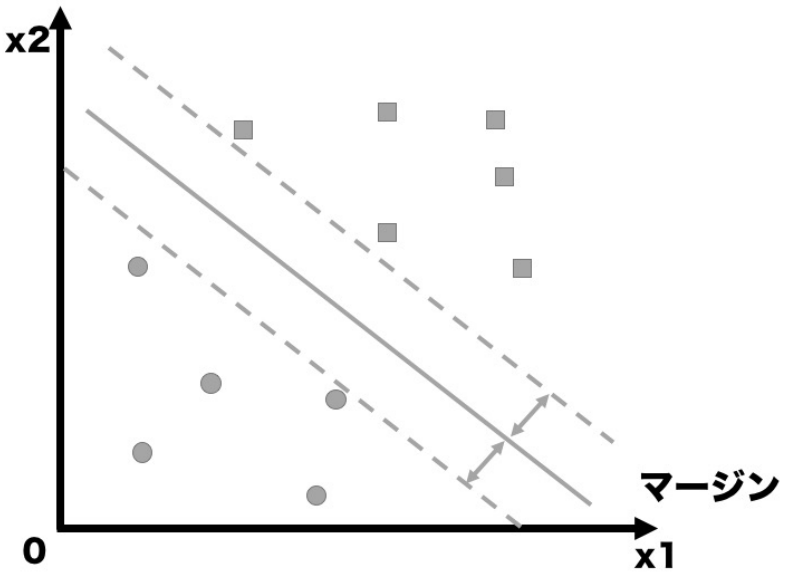

図 2.2　マージン最大化 [4]

2.1.2 単純ベイズ

単純ベイズとはベイズの定理を利用した分類手法であり，迷惑メールの分類などに用いられる．単純ベイズで文書分類を行う際には，文書中の単語の出現頻度を調べ，文書がどのカテゴリに分類するのかを判定する [4]．ナイーブベイズ分類の式は式 2.1 で与えられる [4].

$$P(B|A) = P(B) \times \ P(A|B) \tag{2.1}$$

P(B) は事前確率であり，どのカテゴリに分類させるかの確率を表している．入力テキスト A は，各単語の集合であり，テキストを単語に分割する．また，集合であるため，単語の順序に意味を持たない．入力テキスト A を各単語 aN の集合とすると P(A—B) は式 2.2 になる [4].

$$P(A|B) = P(a_(a1)|B)P(a_(a2)|B)P(a_(a3)|B)...P(a_(xN)|B) \tag{2.2}$$

P(aN|B) の確率は，分割した単語がどのカテゴリに属する確率を求めることである．あるカテゴリにおいて，単語が出現した確率は式 2.3 で求めることが出来る [4].

$$\text{単語の出現率} = \text{単語の出現回数/カテゴリの全単語数} \tag{2.3}$$

学習させる際には出現した単語がカテゴリに分類された回数を保存し，分類を行うさいには P(B) × P(A|B) をカテゴリごとに計算すれば良い．

2.2 既存方式

既存研究では SVM と単純ベイズを用いてツイート分類を行なっている．既存研究では Kleinberg のバースト解析を用いて災害に関連性の強い単語の抽出を行なっている．関連性が強い単語として “水”，“電気”，“ガス”，“放射能” などの単語が挙げられている．これらの単語は震災発生直後に Twitter に出現する頻度が上昇しているが，“電気” や “ガス” に関しては時間とともに出現頻度が現象しており，停電やガスなどの問題が落ち着いてきていることがわかる．この様に重要度の高い単語の出現頻度などで被災地で問題になっていること，必要とされている物を把握することができる．しかし，“水” という単語は飲み水の意

味である「飲料水」や飲み水とは関係ない「御茶ノ水」などの様々な意味で使用されている．そのため，ツイートデータにそのまま分析処理を行なっただけでは，正しい分析結果は得られない．そこで，既存研究では “水” という単語の分類を行なった．“水” が飲み水の意味か，それ以外の意味かで分類することによって，精度の高い分析結果が得られる．分類手法としては SVM 及び単純ベイズを用いて分類を行なっている．その結果，SVM の分類結果は表 2.1 になり，単純ベイズの分類結果は表 2.2 になった．

表 2.1　SVM の分類結果

予測結果 / 実際	飲み水関係あり	飲み水関係なし	合計
飲み水関係あり	14	4	18
飲み水関係なし	7	5	12
合計	21	9	30

表 2.2　単純ベイズの分類結果

予測結果 / 実際	飲み水関係あり	飲み水関係なし	合計
飲み水関係あり	15	3	18
飲み水関係なし	6	6	12
合計	21	9	30

第 3 章

提案手法

既存方式では SVM 及び単純ベイズを用いて分類を行なっている．その分類精度は 70%であった．分類精度が 70%の原因として既存方式では文書分類に必要な特徴を十分に捉えられておらず，情報が欠落している可能性があると考える．そこで，本提案では LSTM を用いて文書分類を行う．LSTM は NN であり，学習時に文書分類に必要な情報を自動的に求めることが可能である．

3.1 手法

3.1.1 RNN

RNN とは再帰型ニューラルネットワークであり，前回の計算を次回の計算に関連付けて学習できる手法である．図 3.1 は RNN の概略図である．データは入力層から中間層，最後に出力層に入り，結果が出力される．ここで前回の中間層の状態を次回に引き継ぐことで前後に依存関係を持たせることが可能となる．また，RNN の学習には確率的勾配降下法が使用され，隠層の重みの誤差を計算する必要がある [5]．誤差を計算する方法として BPTT 法 (backpropagation throughtime) がある．BPTTh 法では図 3.2 の様に最終層から各層の逆伝播を計算し，その後，重みの誤差勾配を計算する [5]．各重みを勾配方向に減少させる勾配降下法の重みの更新を行う．この様に前回と次回を関連づけることが出来るため RNN は時系列データを扱うことが可能となる．自然言語処理においては，文書中の単語の順序を時系列データとして扱うことで処理を可能としている．しかし，RNN には問題点があり，階層が多くなりすぎるとデータの依存関係を保持出来なくなる問題が存在する．この問題は順

伝播型ネットワークの勾配消失問題と同じ原因で起きている [5].

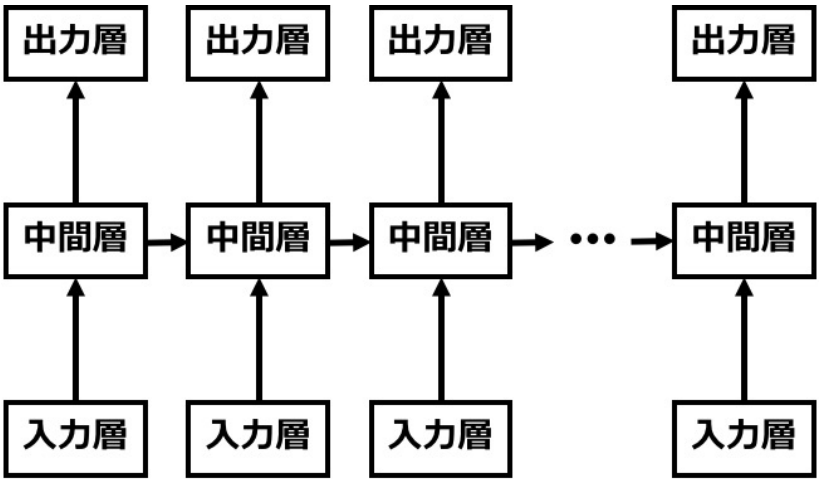

図 3.1　RNN の概要 [6]

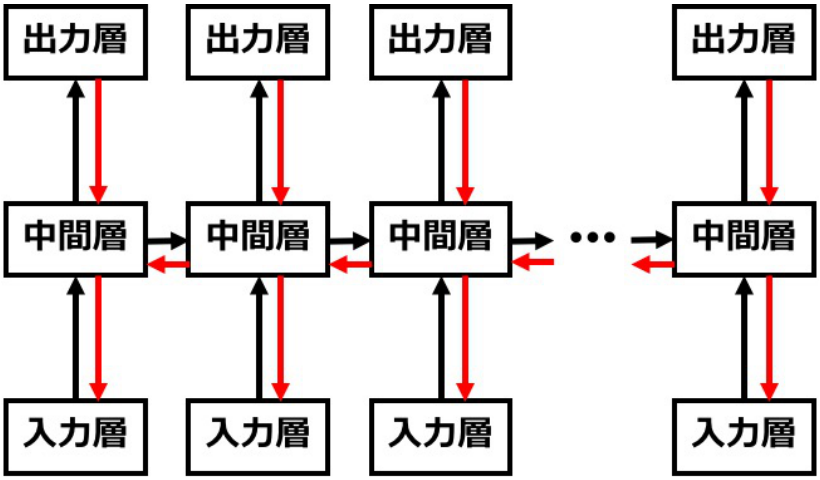

図 3.2　RNN の逆伝播 [6]

3.1.2　LSTM

RNN の一種として LSTM が存在する．LSTM は RNN の階層が多くなると依存関係を保持出来なくなる問題を解決するために作成された．図 3.1 では中間層であった部分が図 3.3 のように LSTM ブロックに置き換わった物が LSTM である．LSTM を用いることで長期的な依存関係を保持出来る様になる．

3.1　手法

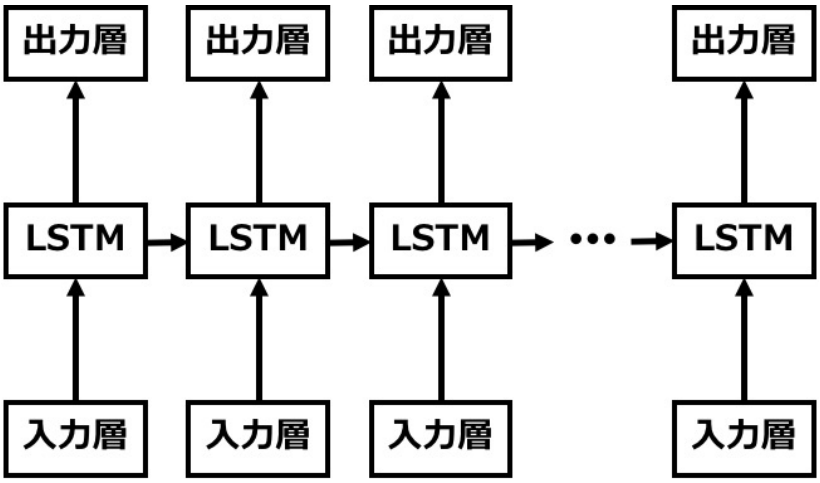

図 3.3　LSTM の概要

第 4 章

実験

4.1　実験環境

実験は以下の環境で行なった．

- OS：Ubuntu 16.04 LTS
- 言語：Python2.7.14
- パッケージ：Chainer 3.3.0
- 形態素解析エンジン：MeCab

4.2　学習アルゴリズム

使用するアルゴリズムとして LSTM を使用する．また，比較のため既存研究でも使用されていた SVM 及び単純ベイズによる分類も行う．

4.3　構成した NN

本実験で構成した NN の詳細は下記のとおりである．

- EmbedID(7176，600)：埋め込み層
- LSTM(600，600)：中間層
- Linear(600，2)：全結合層
- Optimizer：Adam

- Batchsize：81
- Epoch：50

4.4　使用データ

学習させるデータは東日本大震災・熊本自身の発生時である下記の期間から収集したツイートデータ 2400 個を使用する．ツイートは “水” 含みハッシュタグ#jishin，#tsunami や地震に関係のある単語を含んでいるツイートである．また，ツイートデータ 2400 個の中で飲み水に関係のあるツイートが 1200 個，関係のないツイートが 1200 個である．

- 2011 年 3 月 10 日 〜 2011 年 4 月 4 日
- 2016 年 4 月 13 日 〜 2016 年 5 月 13 日

4.5　データの前処理

ツイートデータを学習させる前にデータの前処理を行う．データの前処理は下記の順番に行う．

1. データのラベル付け

   データのラベル付けは学習させるための教師データとして必要である．今回の実験では飲み水に関係のあるツイートにラベル 1 を，関係のないツイートにラベル 2 をラベルとしてラベリングを行なった．
2. 文書から不要な文字の削除

   不要な文字の削除では URL や記号などの分類に不要と思われる文字の削除を行なった．
3. 文書の分かち書き

   分かち書きとは文書を単語ごとに区切り，その間に空白を置く処理のことである．分かち書きは形態素解析エンジン MeCab を使用した．MeCab の辞書は IPA が提供している辞書を用いた．

4. 単語を単語 ID に変換

   単語 ID への変換では学習させるツイートデータから単語と ID の辞書を作成し，辞書にしたがって変換した．また，モデルの使用時に未知の単語が出現するため，未知の単語の ID は 0 に設定した．

以上の処理が学習前の処理となる．

4.6　学習

ツイートデータの学習を行った結果，学習時のロスとテスト時のロスは図 4.1 の様になった．また，学習時のアキュラシーとテスト時のアキュラシーは図 4.2 の様になった．図の epoch は学習回数，loss は分類器の損失率，acc はモデルのアキュラシーを表す．

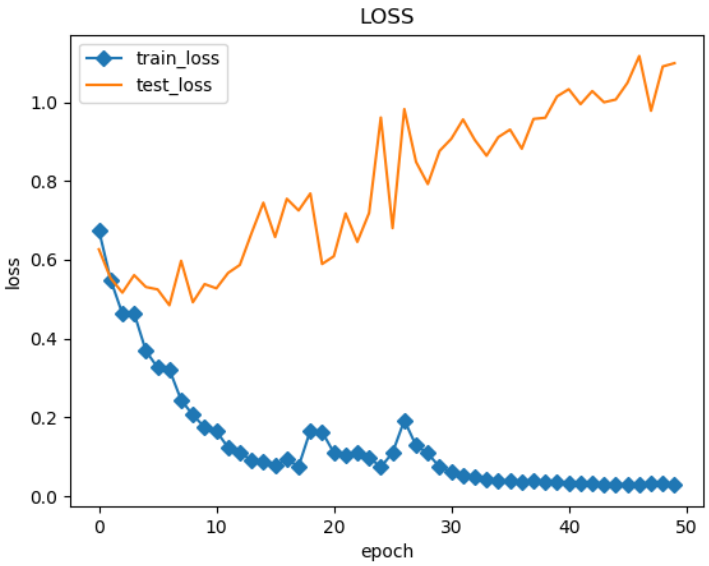

図 4.1　モデルの LOSS

4.6　学習

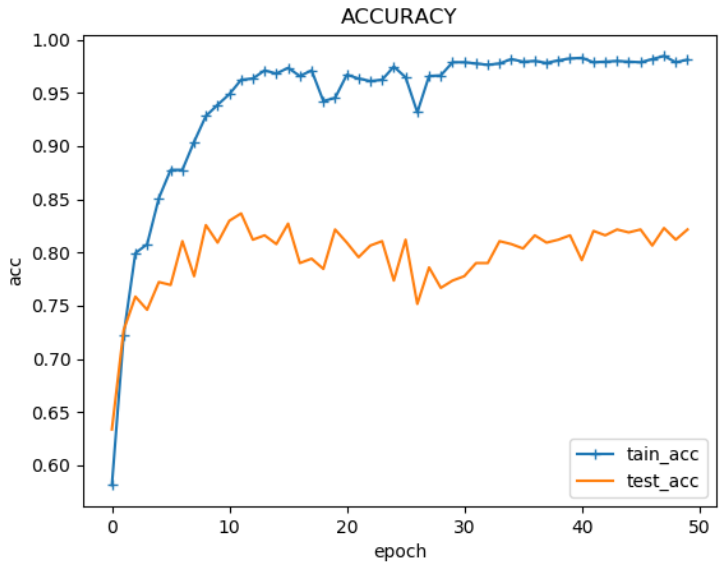

図 4.2　モデルの Accuracy

本実験において学習及び，テストに掛かった時間は表 4.1 の通りである．また，単純ベイズ，SVM の学習に掛かった時間は表 4.2 である．

表 4.1　LSTM の学習・テスト時間

	学習時	テスト時
1epoch の平均	83.3	21.0
合計時間	4163.3	1070.2

表 4.2　単純ベイズ・SVM の学習時間

	単純ベイズ	SVM
学習時間	0.15	11125.16

第 5 章

評価

5.1　モデルの評価

作成したモデルを用いて評価を行なう．モデルの評価には学習時に使用したツイートデータ 2400 個とは別に集めたツイートデータ 200 個を用いる．また，ツイートデータ 200 の中で飲み水に関係のあるツイートが 200 個，関係のないツイートが 200 個である．評価の結果は表 5.1 である．分類において LSTM は分類精度 87%となった．既存方式で使用されていた SVM と単純ベイズにおいても同じ学習データ，評価データを用いて分類を行なった．その結果は表 5.2 になり LSTM が最も分類精度が高いことが確認できた．

表 5.1　モデルの評価

予測結果 / 実際	飲み水関係あり	飲み水関係なし	合計
飲み水関係あり	78	22	100
飲み水関係なし	4	96	100
合計	82	118	200

表 5.2　分類精度の比較

	単純ベイズ	SVM	LSTM
分類精度	84%	72.5%	87%

5.2　結果・考察

実験の結果として，既存研究で用いた手法と比較した際には，SVM 及び LSTM が近い精度になり，単純ベイズの精度が若干下がる結果になった．また，既存研究の結果と比べると，単純ベイズ，SVM ともに分類精度は向上していることがわかった．今回の実験では学習に用いたデータが既存研究より多く用いたため，汎化性能が高い SVM や LSTM において高い精度になったと思われる．実験の結果から LSTM が最も文書分類の特徴を捉えていることがわかるが，SVM も LSTM ほどではないが特徴を捉えることが出来ていることがわかる．今回の実験において作成した分類器は飲み水に関係あると判定したツイートに対して高い精度を誇っていた．しかし，反対に飲み水に関係なしと判定したツイートには約 2 割の間違いがあるため，精度は低いと言える．そのため，飲み水に関係ないツイートに対しての分類精度を向上させる必要がある．

分類出来なかったツイートの特徴は次の様な特徴を持っていた．

1. 一つのツイートに複数回水が出てきている
2. 未知の単語が多い

一つのツイートに飲み水に関係のある水が複数回や，飲み水に関係のある水と関係のない水が含まれていた場合には，判別出来ていないツイートが存在していた．また，未知の単語が出現したツイートの分類に対して，正しく分類出来ていない場合が存在していた．

今回の実験において実験時間は LSTM が役 1 時間 28 分に対して，単純ベイズは約 0.15 秒，SVM が約 3 時間 6 分となった．速度では単純ベイズが圧倒的に早いものの分類精度が低くなっている．近い分類精度の LSTM と SVM では LSTM の方が倍近く早く学習が終了している．分類精度が高く学習時間が早いため，ツイート分類においては LSTM を用いるのが適していると言える．

第 6 章

まとめ

本稿ではツイート分類としてツイートに出現した “水” が飲み水に関係あるか関係ないかの分類を行なった．分類手法として RNN の一種である LSTM を用いて分類を行い SVM，単純ベイズと分類精度を比較した結果，LSTM が最も分類精度が高く，最も分類に必要な特徴を捉えていることを証明した．また，SVM も LSTM ほどではないが特徴を捉えていることがわかった．

今後の課題として，分類精度がまだ 87%であるためより向上させる必要がある．様々な NN の構造を試すことでさらなる分類精度の向上が見込まれる．また，今回分類出来ていなかった，複数回水が出現するツイートを分類するために，複数回水が出現するツイートを学習させる必要がある．また，汎化性能を上げるためにさらに多くの学習データを学習させる必要がある．

謝辞

本研究の遂行と論文作成にあたり，ご指導，ご助言をいただきました高知工科大学情報学群　清水明宏教授に心より感謝申し上げます．また，本研究の副査を担当していただきました高知工科大学情報学群　福本昌弘教授，吉田真一准教授に深く御礼申し上げます．

参考文献

[1] 佐々木智也，“拡大を続ける Twitter の震災における活躍と今後の展望-サービス開始から 5 年、コミュニケーションツールから社会インフラへ-”，http://www.yhmf.jp/pdf/activity/adstudies/vol_36_01_04.pdf，2018 年 2 月 20 日閲覧.

[2] 総務省，“熊本地震と新たな災害情報等の共有の在り方”，http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h29/pdf/n5300000.pdf，2018 年 1 月 27 日閲覧.

[3] 坂巻英一，亀井悦子，“Twitter 上のつぶやきに関するテキストマイニングの事例研究-大規模災害発生時の被災地における現状把握への応用 - ”，日本経営工学会論文誌 65 巻，p.39-50，2014-2015.

[4] クジラ飛行机，“Python によるスクレイピング&機械学習開発テクニック Beautiful-Soup、scikit-learn、TensorFlow を使ってみよう”，2016

[5] 岡谷貴之，“機械学習プロフェッショナルシリーズ深層学習”，2015

[6] 竹田卓也，“リカレントニューラルネットワークの概要と動作原理”，https://wbawakate.jp/data/event/5/rnn.pdf，2018 年 2 月 16 日閲覧