F-019

構成文に基づく災害ツイートのフィルタリング性能の改善
Improving filtering performance of disaster tweets based on individual sentences

小川 智広† 村田 健悟† 伊東 聖矢‡ 大原 剛三‡
Tomohiro Ogawa Kengo Murata Seiya Ito Kouzou Ohara

1 はじめに

マイクロブログは即時性と内容の多角性の面から災害時の情報源として有効と考えられており，実際に Twitter[1] 上のツイートから災害時に有益な情報が取得できることが示されている [8]．災害時には膨大な量のツイートが投稿され，人手による情報収集が困難になるため，一般にはキーワードを指定してツイートを取得する方法が用いられる．しかし，キーワードを指定して取得したツイートには災害に関係のないツイートが含まれるため，災害に関連する有益なツイートとそれ以外を分類する研究が行われている [1, 2, 4, 6, 7, 9]．これらのツイートを対象とした災害情報分類は，災害時という緊急性の高い状況下での利用を想定している．さらに，災害時に有益な情報を含むツイートは，キーワード指定によって取得した全ツイートの中にごく少数しか含まれていないことが想定される．そのため，災害時に有益な情報を含まないツイートを有益であると誤分類することを防ぐと同時に，災害時に有益な情報を含むツイートを短時間で取りこぼしのないように特定することが求められる．

これまでの災害情報分類手法の多くは，Support Vector Machine（SVM），単純ベイズを始めとする従来の機械学習手法のほか，畳み込みニューラルネットワーク（CNN: Convolutional Neural Network），再帰型ニューラルネットワーク（RNN: Recurrent Neural Network），Bidirectional Encoder Representations from Transformers（BERT） [3] などの深層学習手法を用いている．特に，BERT を用いた手法は，従来の機械学習手法や他の深層学習手法よりも高い分類精度を示すことが報告されている [5]．泉ら [9] の研究では，日本国内で観測された台風と大雨時に投稿されたツイートから取得した「台風」と「雨」を含むツイートに災害時に必要か否かのラベルを付けたツイートデータを利用し，BERT モデルの評価をしている．結果として，正解率が 81.2%，真陽性率が 76.0% となっている．さらに，泉らはツイートに含まれる一般的でない文章構造に誤分類の要因があると考察している．一方，災害時に必要なツイート数と不必要なツイート数がほぼ同数となっている泉らのテストデータとは異なり，実際の分布に従ったテストデータを用いた我々による評価では，誤って必要と推論されるツイートの割合が多くなることで F 値が低下することを確認した．

このような背景の下，本研究では，災害発生時にキーワードを指定して取得したツイート群から災害時に有益なツイートを分類するタスクにおいて，ツイートの誤分類を軽減することを目指す．そのために，提案手法では，話題の異なる複数の文から構成されるツイートが誤分類されやすいという予備実験の結果に着目し，ツイートを文単位に分割し，分割後の文に対して既存の深層学習手法を用いることでツイートの誤分類を軽減することを試みる．一方，ツイートを文単位に分割した場合，災害時に有益でない文がもつ話題は多岐にわたり，その数には偏りが生じ，その結果，学習データ数が少ない話題に関する文は学習が不足し，さらなる誤分類の要因になる可能性がある．この不均衡問題に対し，提案手法では，まず，少数しか存在しない災害時に有益な情報を含む文と多数存在する有益でない文を均衡化し，さらに文の話題を推定するトピックモデルを用いて話題ごとの文数の均衡化を図る．文単位への分割に加えて，これらの均衡化を行うことで，ツイートに深層学習手法を直接適用する場合と比較して，災害関連ツイートに対する誤分類を軽減させる．評価実験では，実際の日本語の災害関連ツイートを用いて提案手法の分類性能を定量的に評価し，既存手法と比べて F 値を改善できることを確認する．

2 提案手法

2.1 災害関連ツイート・文の定義

本研究におけるツイートを構成する文に対する教師ラベルは，災害関連情報として必要か不必要かの 2 値であり，必要・不必要の定義は，Parilla-Ferrer ら [7] の定義を基に定めた．必要ラベルが付与されるのは，対象災害や対象災害による被害の状況把握に有益な文であり，不必要ラベルが付与されるのは，具体的な状況が掴めない文，オフトピックな文，噂・ジョーク・スパム・荒らし・災害と関連しない内容を含む文とした．また，ツイート単位でのラベルを付与する際は，必要ラベルが付与された文が 1 文でも含まれるツイートは必要ツイートとした．具体的には，「台風の雨と風がめっちゃ強い！木も倒れてたし、まだ家にいるか。もうこの際アニメ一気見するか笑」のようなツイートは，1 文目で台風の雨と風が強いという対象災害の状況，2 文目で木が倒れているという対象災害による被害の状況が把握できるため，必要ツイートと定義する．また，「今日の予定めんどすぎる。台風でもくればいいのに！」のようなツイートは，2 文目で台風について言及しているが，対象災害の状況やその被害状況の把握はできないため，不必要ツイートと定義する．

2.2 提案手法の概要

本研究では，災害関連ツイートの分類タスクにおいて，ツイート本文を分割して文の集合としたうえで，各文に付与されているラベルを教師データとして用いる．さらに，トピックモデルによって付与されたトピックごとの学習データ数が均等になるようにすることで，災害関連情報として必要なツイートに対する F 値の改善を試みる．図 1 に提案手法の概要図を示す．

† 青山学院大学大学院理工学研究科
‡ 青山学院大学理工学部
1 https://twitter.com

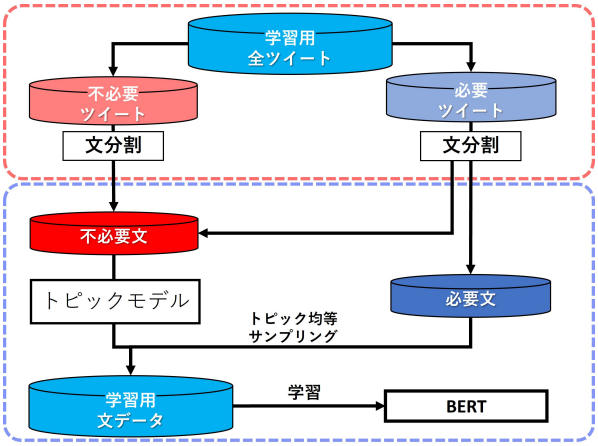

図 1　提案手法におけるファインチューニングの概要図

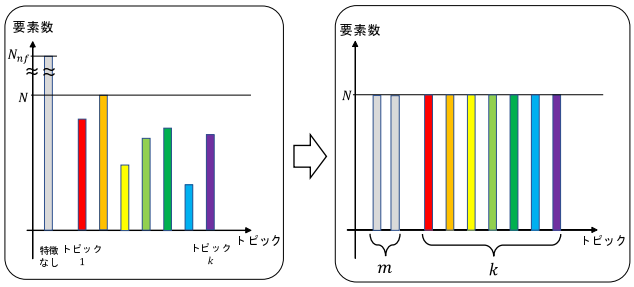

図 2　不必要文の補完方法

2.3　トピックモデルを用いた均衡サンプリング

学習用ツイート集合では，元々不必要ツイートの方が多いため，ツイートを文単位に分割するのみでは不必要文の数が大幅に増加する．さらに，不必要文に対する分析結果から，不必要文の話題のバリエーションが極端に多いことに加え，各話題に属する不必要文の数が大きく異なることが判明している．これに対して，必要文・不必要文のデータ数が均等になるように必要文を単にオーバーサンプリングするだけでは，学習データ数が少ないトピックに属する不必要文の学習が不足し，誤分類の要因となり得る．そのため，出現頻度が少なく学習が不足している話題に対する不必要文の学習データを補完する必要がある．

そこで提案手法では，不必要文に対して文の話題を考慮した均等サンプリングを行う．具体的には，学習データ中の各不必要文に対して，トピック数を k に設定したトピックモデルを用いてトピックの構成割合を算出し，構成割合が最大となるトピックをその不必要文におけるトピックとする．この時，ストップワードの除去によりトピックが判定不能となった文，すべてのトピックに対する割合が同一となった文は特徴なしトピックとして取り扱う．これ以降は，各不必要文に付与されたトピックを用いて学習データの補完を行う．まず，図 2 に示すように，各トピックに対する不必要文を，全トピックにおける最大要素数 N 件分復元抽出する．特徴なしトピックは，特徴なしトピックの要素数を N_{nf} としたとき，$N_{nf} \leqq mN$ を満たす最小の自然数 m に対して mN 件分復元抽出する．これにより，$(k+m) \times N$ 件分の不必要文がサンプリングされる． 最後に，図 3 に示すように，必要文と不必要文に関して，バッチサイズを B として

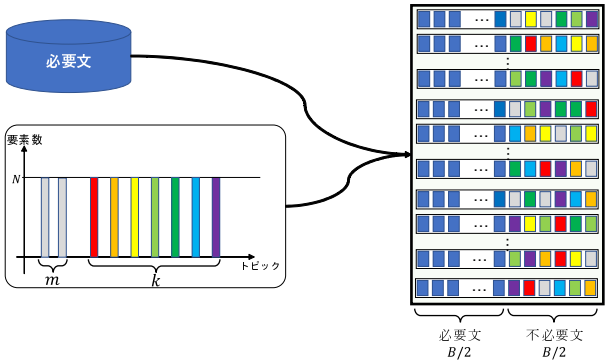

図 3　学習データ作成方法

表 1　ツイートラベルの分布

	必要ツイート	不必要ツイート
学習用データ	304 件	1,957 件
テスト用データ	68 件	932 件

必要文を $B/2$ 件復元抽出し，不必要文を $B/2$ 件非復元抽出することにより，教師ラベルの分布，および，不必要文のトピックの分布が均衡化された学習データを作成する．

3　評価実験

3.1　データセット

本研究では，日本語の災害関連ツイートの分類を対象とするため，日本語ツイートに対して災害時に必要か否かの教師ラベルが付与（アノテーション）されたデータセットが必要となる．しかし，そのようなラベルがアノテーションされた日本語の災害関連ツイートデータセットは公開されていない．そこで本研究では，泉ら [9] と同じ条件で Twitter API[2]を用いて収集した日本語ツイートに対して，1 ツイートあたり 5 人のアノテータによるアノテーションを実施することでデータセットを作成した．具体的には，学習データ 2,261 件，テストデータ 1,000 件となるデータセットを作成し，泉らの手法 [9] と提案手法を評価した．複数のデータセットでの評価実験を行うため，学習データ，テストデータの件数と教師ラベルの割合を維持して無作為にツイートを入れ替えることで，20 個のデータセットを作成した． なお，1 ツイート当たりの平均文数は 2.63 文，最大文数は 19 文，最小文数は 1 文である．

3.2　実験設定

3.2.1　比較手法

本実験では，ツイート単位で学習・推論する手法（比較手法）と文単位で学習・推論する提案手法の分類性能を定量的に評価した．比較手法には，泉ら [9] と同様に事前学習済み BERT モデルを用い，ツイート単位の学習データでファインチューニングを行った．ただし，本実験で用いたデータセットが不必要ツイートと比較して必要ツイートが極端に少ない不均衡なものであることを考慮し，ミニバッチ内均衡サンプリングの処理を追加している．具体的には，必要ツイートは復元抽出，不必要ツイートは非復元抽出を行うことで学習時の教師ラベルの

2　https://developer.twitter.com/ja/docs/twitter-api

表 2　各手法の学習設定

手法	エポック数	学習率	バッチサイズ
比較手法	18	5×10^{-7}	16
提案手法	17	2×10^{-7}	16
文ラベル手法	17	3×10^{-7}	32

表 3　比較手法と提案手法の分類精度

手法	真陽性率	真陰性率	適合率	F 値
比較手法	**0.811**	0.881	0.332	0.471
提案手法	0.794	**0.899**	**0.367**	**0.501**

偏りを抑制した．提案手法においても同じ事前学習済み BERT モデルを利用し，前述のトピックモデルを用いた均衡サンプリングを行った文単位の学習データを用いてファインチューニングした．一方，テストデータに対する予測は，比較手法との対比のためにツイート単位とした．具体的には，ツイート中の各文に対する推論結果の論理和を取り，1 文でも必要と推論された文があればそのツイートに対する推論結果は必要とし，それ以外の場合は不必要とした．また，提案手法におけるトピックモデルを用いた均衡サンプリングの効果を確認するために，ミニバッチ内均衡サンプリングの処理を行った文単位の学習データを用いてファインチューニングした手法（文ラベル手法）とも比較した．文ラベル手法のテストデータに対する予測は，提案手法と同様に，ツイート中の各文の推論結果の論理和によるツイート単位で行い，1 文でも必要と推論された文があればツイート単位での推論結果は必要とし，それ以外の場合は不必要とした．さらに，提案手法と文ラベル手法に関しては，テストデータにおける文単位の分類性能も比較した．

3.2.2　学習設定

本実験で利用する BERT では，エポック数，バッチサイズ，学習率を除く他の学習設定は，google-research[3] が公開する BERT モデルのデフォルト値を利用した．エポック数，バッチサイズ，学習率は，比較手法がツイート単位，提案手法が文単位とデータの粒度が異なることを考慮し，適宜調整することで表 2 のように設定した． また，トピックモデルには，Python ライブラリの Gensim が提供する LDA モデル[4]を利用し，トピック数を 15 に設定した．トピックの均等サンプリングを行う際に使用する最大要素数 N は，複数のデータセットを作成した場合のバラつきを解消するため，データセットごとの最大要素数 N の平均値 1,050 を用いた．

3.3　定量評価

比較手法と提案手法の分類精度を真陽性率，真陰性率，適合率，F 値によって評価した結果を表 3 に示す．また，表 4 と表 5 は，比較手法と提案手法におけるツイート単位での分類結果を混同行列で表したものである．なお，本稿における実験結果は，比較手法と提案手法はいずれも 20 試行，文ラベル手法は 10 試行に対する平均値である． 表 3 に示すように，比較手法に比べて提案手法は，適合率が約 3.5 ポイント改善し，F 値が約 3.0 ポイント改善した． さらに，表 4 と表 5 に示すように，提案手法は比較手法と比べて，真陽性ツイート数が約 1 件減少しているのに対し，偽陽性ツイート数が約 17.4 件減少していることから，総合的な分類精度が改善していることがわかる．また，真陽性率が約 1.7 ポイント低下している点については，不必要文と必要文のサンプル数の差が顕著であり，それに対する不均衡対策の効果が薄かったことが原因と考えられる．

これらの結果から，提案手法は比較手法と比べて，必要ツイートに対する分類性能をある程度維持し，同時に，不必要ツイートに対する誤分類を抑制することが可能になったと考えられる．

3 https://github.com/google-research/bert
4 https://radimrehurek.com/gensim/models/ldamodel.html

表 4　比較手法の混同行列

		予測ラベル	
		必要	不必要
正解ラベル	必要	55.1	12.9
	不必要	111.3	820.7

表 5　提案手法の混同行列

		予測ラベル	
		必要	不必要
正解ラベル	必要	54.0	14.0
	不必要	93.9	838.1

表 6　比較手法と文ラベル手法の分類精度

手法	真陽性率	真陰性率	適合率	F 値
文ラベル手法	**0.875**	0.807	0.250	0.388
比較手法	0.811	**0.881**	**0.332**	**0.471**

表 7　不正解ツイートにおける単文/複文ツイートの件数

	偽陽性	偽陰性
単文ツイート	25 件	4 件
複文ツイート	80 件	10 件

3.4　詳細分析

3.4.1　ツイートラベルと文ラベルでの分類性能比較

提案手法のように文単位の学習データを用いることの効果を検証するために，比較手法と文ラベル手法の分類性能を比較した．ここでのテストデータに対する予測は，ツイート単位とした．各手法に対する評価結果を表 6 に示す． 表 6 より，両手法の分類精度を比較すると，文単位の学習データを用いた場合は，真陽性率は改善しているが，他の評価指標の値は大幅に低下していることがわかる．具体的には，真陰性率が低下していることから，特別な工夫をせずに，単に文単位の学習データを用いて学習し，ツイート単位の推論を行った場合，不必要文を正しく不必要と予測できないために偽陽性のツイートが増加し，その結果，適合率と F 値が悪化したと考えられる．

さらに，比較手法での誤分類の傾向，および文単位の推論が正しくできた場合に期待される分類性能の改善度合いを調査するため，比較手法での不正解ツイートにおける単文ツイート，複文ツイートそれぞれの件数を確認した．その結果を表 7 に示す．表 7 より，比較手法による不正解ツイート中の複文ツイートの割合が 75.6% に上ることがわかる． また，データセット内の全ツイート中の複文ツイートの割合は 66.4% であった．これらのこ

表 8　文ラベル手法と提案手法の分類精度

手法	真陽性率	真陰性率	適合率	F 値
文ラベル手法	**0.875**	0.807	0.250	0.388
提案手法	0.794	**0.899**	**0.367**	**0.501**

表 9　文ラベル手法と提案手法の文単位での分類精度

手法	真陽性率	真陰性率	適合率	F 値
文ラベル手法	**0.825**	0.912	0.227	0.356
提案手法	0.737	**0.950**	**0.342**	**0.466**

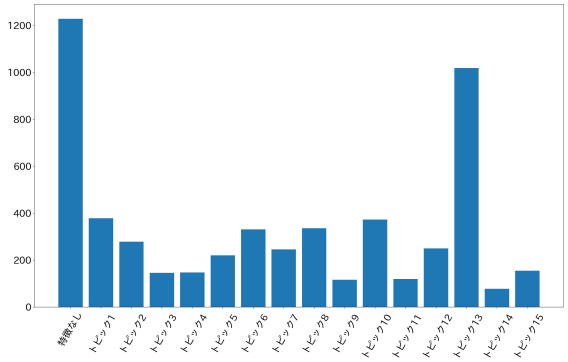

図 4　各トピックの不必要文の要素数の分布

とから，不正解であった複文ツイートの割合が全データセットにおける複文ツイートの割合より大きく，そのような不正解となった複文ツイートに含まれる個々の文に対する分類精度を向上させることができれば，ツイート単位での大幅な精度向上につながることがわかる．

3.4.2　トピックモデルの有無による分類性能比較

次に，文単位で学習する提案手法におけるトピックモデルを用いた不必要文の均等サンプリングの効果を検証するため，文ラベル手法と提案手法のツイート単位の分類精度を比較した．その結果を表 8 に示す．表 8 より，真陽性率を除く他の評価指標すべてにおいて，提案手法の方が分類精度を改善できていることがわかる．

また，両手法の文単位の分類精度を確認した．その結果を表 9 に示す．表 9 からは，文ラベル手法では F 値と適合率が低くなっているのに対し，提案手法ではそれらの値が比較的高くなっていることがわかる． これは，提案手法の真陽性ツイート数が文ラベル手法に比べて低くなっているものの，真陰性ツイート数が多くなっていることから，結果的に偽陽性ツイート数が大幅に減ったためであると考えられる．これらの結果から，文単位のデータでの学習時に不必要文のトピック不均衡問題を解消することが，偽陽性文増加の抑制に寄与していると言える．ここで，学習データにおける不必要文に実際にどれだけの不均衡が生じていたのかを確認した． 不必要文集合にトピックモデルを適用した結果得られた各トピックに対する不必要文数の分布を図 4 に示す．図 4 より，多くのトピックが 200 前後の不必要文をもつものの，特徴なしトピックとトピック 13 ではその数は 1,000 を超え，トピック 14 では 100 未満となっており，大きな不均衡が生じていることがわかる．このような不均衡が提案手法では解消されるため，不必要文がより適切に学習されることにより，不必要文が必要文に誤分類されることが減少したものと考えられる．

4　おわりに

本研究では，災害関連ツイートの分類において，ツイートを文単位に分割し，不必要文に対するトピックの分布を均等にしたうえで学習する分類モデルを提案した．ツイートを文単位に分割することで，複数の話題をもつ文から成るツイートに対する分類性能の改善を可能としている．また，評価実験を通して，ツイート単位で学習・推論する既存アプローチと比較して，提案手法が必要ツイートの分類性能を同程度に保ちながら，不必要ツイートの誤分類を抑制できていることを確認した．今後の課題として，真陽性率の改善のために，必要文の量的な追加だけでなく，それに伴う話題の分布の変化を考慮した学習手法の検討が必要である．

参考文献

[1] Cornelia Caragea, Adrian Silvescu, and Andrea H. Tapia. Identifying informative messages in disaster events using convolutional neural networks. In *Proceedings of the 13th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, pages 137–147, 2016.

[2] Jishnu Ray Chowdhury, Cornelia Caragea, and Doina Caragea. Cross-lingual disaster-related multi-label tweet classification with manifold mixup. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop (ACL)*, pages 292–298, 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186, 2019.

[4] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, pages 1638–1643, 2016.

[5] Anh Duc Le. Disaster tweets classification using bert-based language model. *arXiv preprint arXiv:2202.00792*, 2022.

[6] Venkata Kishore Neppalli, Cornelia Caragea, and Doina Caragea. Deep neural networks versus naive bayes classifiers for identifying informative tweets during disasters. In *Proceedings of the 15th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, pages 677–686, 2018.

[7] Beverly Estephany Parilla-Ferrer, Proceso Fernandez, and Jaime T. Ballena. Automatic classification of disaster-related tweets. In *Proceedings of the International conference on Innovative Engineering Technologies (ICIET)*, pages 62–69, 2015.

[8] 坂巻英一 and 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(1):39–50, 2014.

[9] 泉 翔太, 堀 太成, 山根 達郎, 全 邦釘, 藤森 祥文, and 森脇亮. Deep learning を用いたマイクロブログ投稿文の災害情報分類. *AI ・データサイエンス論文集*, 1(J1):398–405, 2020.