コロナ禍関連ツイートの自動話題分類

山下 竜也 † 熊本 忠彦 ‡

†, ‡ 千葉工業大学 情報科学部 情報ネットワーク学科

1. はじめに

2020 年 1 月以降に拡がった新型コロナウイルス感染症の影響により，不要不急な行動や三密（密閉・密集・密接）が制限され，抑圧された社会生活を過ごさなくてはならなくなったこともあり，このコロナ禍の現状に関する不満や想い，考えを Twitter をはじめとする SNS 上で発信している人が増えている．

そこで本稿では，コロナ禍に関し，人々がどのようなことを感じ，発信しているのかを分析するための手段として，コロナ禍に関連するツイートを話題別に分類する手法を提案するとともに，代表的なトピックモデルの一つである LDA（Latent Dirichlet Allocation）[1]を用いた手法と精度を比較することで，その有用性を検証する．

2. 関連研究

圓谷らは，Amazon レビュー等からなる日本語文章を，BERT [2]を用いてベクトル化し，クラスタリングする手法[3]を提案している．しかしながら，各クラスタへのトピックラベルの付与は手動であり，精度評価もしていない．

3. コロナ禍関連ツイートの取得

まず，一定以上の頻度で Twitter にツイートを投稿しているユーザ 12,326 人から 20 代～50 代の男女をランダムに抽出し，「新型コロナウイルス感染症あるいはコロナ禍の現状」に関するツイートを 1～2 件創作してもらった．その結果，男性 2,202 人，女性 2,144 人の計 4,346 人から 7,538 件のツイートを得ることができたが，無意味な文字列やコロナ禍と全く関係のないツイートなど趣旨に合わないツイートがあったため，削除したところ，最終的には 4,615 件になった．

次に，この 4,615 件を均等に分け，5 つのデータセットを構築し，各データセット内のツイート 923 件を 9 人の作業者に読んでもらい，新型コロナウイルス感染症もしくはコロナ禍の現状に関して書かれているかどうかを「そう思う(3)，どちらかと言えばそう思う(2)，どちらかと言えばそう思わない(1)，全くそう思わない(0)」の 4 段階で評価してもらった．

本稿では，ツイート毎に求めた評価結果の平均が 2.5 以上であった 2,960 件をコロナ禍関連ツイートと定義した．

4. 話題分類手法の提案

本節では，提案手法の流れを示す．

手順 ① LDA による話題語の抽出

LDA を用いて 3 節で得たコロナ禍関連ツイート 2,960 件から 4 つのトピックを抽出した（トピック数はトピックモデルの評価指標である Perplexity と Coherence により著者らが決定した）．このとき，それぞれのトピック毎にトピック確率が高かった上位 5 単語をそのトピックを表す話題語として抽出した．結果を表 1 にまとめるとともに，Python の WordCloud ライブラリを用いて可視化した結果を図1に示す．なお，LDA 実行時に処理対象となる単語を名詞，動詞，形容詞，副詞，接頭詞に限定するとともに，名詞の「コロナ」と動詞の「する」を除外した．

表 1. LDA の結果（話題語とトピック確率）

トピック 0		トピック 1	
早く	0.018	事態	0.050
ワクチン	0.016	緊急	0.049
接種	0.014	宣言	0.049
なっ	0.013	感染	0.041
旅行	0.008	解除	0.032

トピック 2		トピック 3	
早く	0.040	ワクチン	0.023
人	0.024	者	0.020
マスク	0.023	感染	0.019
宣言	0.016	早く	0.015
解除	0.016	新型	0.012

(a) トピック0　(b) トピック1　(c) トピック2　(d) トピック3

図 1. LDA 結果の WordCloud による可視化

手順 ② ツイートベクトルの生成

3 節で得たツイート 2,960 件を fastText[4]を用いてベクトル化した．fastText は 2016 年に Facebook 社（現，メタ社）が公開した自然言語処理ライブラリであり，単語を 300 次元のベクトルに変換することができる．

このツイートのベクトル化では，それぞれのツイートから名詞（「コロナ」を除く），動詞（「する」を除く），形容詞，副詞，接頭詞を抽出し，各単語をベクトル化した後，平均ベクトルを求め，そのツイートのツイートベクトルとした．

手順 ③ トピックベクトルの生成

トピックのベクトル化でも fastText を用いた．すなわち，それぞれのトピック毎に抽出された 5 つの話題語を fastText を用いてベクトル化し，トピック確率を重みとする重み付き平均ベクトルを求めることで，そのトピックのトピックベクトルとした．

手順 ④ コサイン類似度に基づくツイートの話題分類

各ツイートベクトルと全トピックベクトルの間でコサイン類似度を求め，コサイン類似度が最も大きかったトピックをそのツイートのトピックとした．

ここで，各トピックに分類されたツイートの例として，それぞれのトピックベクトルに対し，コサイン類似度が最も大きかった上位 2 件のツイートを表 2 に示す．

Automatic Topic Classification of COVID-19 Related Tweets

† Ryuya Yamashita, Chiba Institute of Technology

‡ Tadahiko Kumamoto, Chiba Institute of Technology

表 2. 各トピックに分類されたツイートの例

トピック	類似度	ツイート
0	0.892	ワクチン接種早く始まるといいな
	0.877	ワクチン接種の遅い JAPAN
1	0.950	緊急事態宣言解除うれしい
	0.948	ワクチンで緊急事態宣言解除
2	0.819	緊急時宣言早く解除してほしい
	0.800	緊急事態宣言の解除はまだ早いと思う。
3	0.823	感染者増加
	0.807	コロナウィルス感染が怖い

5. 精度評価

5.1 正解データの作成

話題分類の正解データを作成するために，3 節で得た 2,960 件のコロナ禍関連ツイートを読み，図1に示したような LDA の結果を可視化した図を見ながら，それぞれのツイートがどのトピック（複数可）に該当するかを判定するという作業を 5 人の作業者に行ってもらった．

この作業結果を集計し，それぞれのツイートにおいて投票数の最も多かったトピック（複数可）をそのツイートのトピック（正解データ）とした．結果，3 つのトピックが選ばれたツイートが 7 件，2 つが 293 件，1 つが 2,660 件であった．一方，5 人全員の判定が一致したツイートは 680 件，4 人一致は 893 人，3 人一致は 1,077 人，2 人一致は 310 件であった．人にとっても分類するのが難しいツイートが一定数あったことがわかる．

5.2 LDA を用いた手法による話題分類

4.1 節において LDA により話題語を抽出した際，各ツイートにもトピック毎のトピック確率が算出されている．そこで，トピック確率が最も高かったトピック（複数可）をそのツイートのトピックとした．結果，トピック数が 4 つ全部になったツイートが 15 件，2 つになったツイートが 8 件あり，残り 2,937 件は 1 つであった．

5.3 提案手法と LDA を用いた手法の精度比較

提案手法と LDA を用いた手法の精度を比較するために，それぞれの手法による話題分類の結果と正解データから混同行列を作成し，トピック毎の適合率・再現率・F1 値とそれぞれのマクロ平均を求めた．2 つの混同行列とトピック毎の精度をそれぞれ表 3 と表 4 に示す．

表 4 より，全体的な精度（マクロ平均 F1 値）という意味では，提案手法と LDA 手法はほぼ同等といえるが，トピック毎の F1 値は LDA 手法の方が高いことがわかる．ここで，表 3 の混同行列を見てみると，両手法とも多くのツイートをトピック 2 に分類しており，精度低下の一因となっている．クラスタ間の違いをより明確に表現できれば高精度化も可能と考えられる．また，表 3 に示した 2 つの混同行列を見比べてみると，分布傾向が異なることがわかる．この分布傾向の違いを利用して，お互いの手法を相補的に用いることができればより一層の高精度化もできる可能性がある．

6. まとめ

本稿では，コロナ禍に関し，人々がどのようなことを感じ，発信しているのかを分析するための手段として，コロナ禍に関連するツイートを話題別に分類する手法を提案した．提案手法は，2,960 件のコロナ禍関連ツイートに対し，LDA[1]によるトピック（5 つの話題語）の抽出，fastText[4]によるツイートとトピックのベクトル化，コサイン類似度に基づくツイートの話題分類を行うことで，各ツイートを 4 つのトピックに分類することができた．LDA を用いた手法と精度を比較したところ，ほぼ同等の精度を得ることができた．

今後の課題としては，任意のツイートに対するコロナ禍関連度の自動判定やトピック数の自動決定，ツイートの話題分類基準の高精度化，コロナ禍関連ツイート以外への適用拡張などが挙げられる．

表 3. 混同行列

		トピック 0	トピック 1	トピック 2	トピック 3
		提案手法による話題分類の結果			
正解データ	0	247	0	171	0
	1	65	153	919	19
	2	26	12	780	10
	3	168	7	570	120
		LDA 手法による話題分類の結果			
正解データ	0	193	25	95	107
	1	213	402	376	171
	2	165	75	494	115
	3	173	290	176	256

表 4. 適合率・再現率・F1 値と各マクロ平均

トピック		0	1	2	3	マクロ平均
提案手法	適合率	48.8%	89.0%	32.0%	80.5%	62.6%
	再現率	59.1%	13.2%	94.2%	13.9%	45.1%
	F1 値	53.5%	23.0%	47.7%	23.7%	37.0%
LDA 手法	適合率	25.9%	50.8%	43.3%	39.4%	39.9%
	再現率	46.0%	34.6%	58.2%	28.6%	41.8%
	F1 値	33.2%	41.1%	49.6%	33.2%	39.3%

謝辞

本研究の一部は，JSPS 科研費 20K12085 により実施されている．ここに記して深く感謝の意を表す．

参考文献

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

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” arXiv:1810.04805v2, 2018.

[3] 圓谷顯信，高橋宏和，安達由洋，BERT による日本語文の感情分析と話題分析，第 84 回情処全大，4C-06，2022.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov, “Enriching Word Vectors with Subword Information,” Transactions of the Association for Computational Linguistics, Vol. 5, pp. 135-146, 2017.