CD-002

メトリックの統合によるツイート話題分類の高精度化に関する検討

A Study on Improving Performance of Tweet Topic Classification by Integrating Metrics

熊本 忠彦 †
Tadahiko Kumamoto†

1. はじめに

近年，情報通信インフラの発展と普及に伴い，Social Networking Service（SNS）を利用する人が増えている [1]．特に，マイクロブログの一つである Twitter[2] は，他の SNS と比べ，一度に投稿できる文字数が 140 字までと比較的少なく，自分の本当のプロフィールを明らかにする必要もないことから，年齢や性別，職業といった立場を考えずに気軽に投稿できるという特徴がある．そのため，Twitter では，日々の身近なことから社会全体に対することまで様々な話題に関して多種多様な意見やコメント等が発信されている．

様々な話題の中から特定の話題に関するツイートを集め，より粒度の細かい話題に分類することで，その話題に関し，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ことが可能となる．例えば，2020 年以降に起きた新型コロナウイルス感染症の世界的流行では，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，新型コロナウイルス感染症やコロナ禍の現状などに関する不平や不満，想い，考えが数多くツイートされている．こういったツイートに話題分類を適用することで，コロナ禍に関する多くの人々の本音を分析することが可能となり，QOL（生活の質 ; Quality Of Life）の向上に資するものと考えられる．

そこで本稿では，コロナ禍に関連するツイートを対象に，より粒度の細かい話題に分類する深層学習ベースの手法 [3] を示し，その高精度化について検討する．具体的には，ツイートとそのツイートから抽出されるトピックをベクトル化し，両者を比較することで，より高精度な話題分類を実現する．トピックの抽出には，代表的なトピックモデルの一つである Latent Dirichlet Allocation（LDA）[4] を用い，ツイートやトピックのベクトル化には，深層学習ベースの分散表現生成手法の一つである fastText[5] を用いる．なお，ベクトルどうしの近さを測るためのメトリックとして，一般的にはコサイン類似度かユークリッド距離のいずれかを用いることが多いが，本稿ではこの 2 つのメトリックと LDA が算出するトピック確率（各ツイートがそれぞれのトピックに属する確率）を統合することで，より高精度な話題分類を実現することができることを示す．

以下，本稿では，**2** 章で関連研究を示し，本研究の新規性について述べる．**3** 章で話題分類の対象となるコロナ禍関連ツイートを定義し，取得する．**4** 章でツイートを話題分類するための深層学習ベースの手法 [3] について述べる．**5** 章でコサイン類似度，ユークリッド距離，トピック確率という 3 つのメトリックを統合することで，ベースライン手法（LDA のみを用いた手法）より高精度な話題分類を実現することができることを示す．最後に，**6** 章で本稿のまとめと今後の課題について述べる．

2. 関連研究

圓谷らは，日本語文章をベクトル化し，クラスタリングすることで，話題分類する手法 [6][7] をいくつか提案している．例えば，文献 [6] では，Amazon レビューや Google クチコミ等からなる日本語文章を深層学習ベースの分散表現生成手法の一つである BERT[8] を用いてベクトル化し，非階層型クラスタリング手法の一つである k-means++法 [9] を用いてクラスタリングしている．一方，文献 [7] では，Livedoor ニュースコーパス等からなる日本語文章を文章向けの BERT とも言える Sentence-BERT[10] を用いてベクトル化し，階層型クラスタリング手法の一つである Ward 法を用いてクラスタリングしている．しかしながら，いずれの手法も粒度の細かい話題分類を目指したものではなく，比較的粗い分類を行っている．また，クラスタリングの際に用いたメトリックについての言及はないが，k-means++法や Ward 法ではユークリッド距離を用いることが多い．コサイン距離（= 1 − コサイン類似度）やマハラノビス距離などの使用も考えられるが，いずれにせよ用いられるメトリックは 1 種類のみであり，メトリックの統合は考えていない．

Zhang らは，10〜20 words からなる短い文章を話題分類するためのトピックモデルとして FastText-based Sentence-LDA モデル [11] を提案している．単語ベクトルどうしの近さを測るためのメトリックにはコサイン距離を用いている．

西田らや王らは，Twitter に投稿された膨大なツイートの中から特定の話題に関するツイートを抽出するために，ツイートの圧縮され易さを応用した手法 [12][13] を提案している．しかしながら，これらの手法は，多分類ではなく二値分類（特定の話題か否か）を行うものであり，粒度の細かい話題分類には対応していない．また，

† 千葉工業大学 情報科学部 情報ネットワーク学科
Chiba Institute of Technology,
2–17–1, Tsudanuma, Narashino, 275–0016 Chiba, Japan

メトリックにはデータ圧縮後のデータサイズの変化量が用いられている．

いずれの先行研究も複数のメトリックの統合を考えていない点が本研究とは異なっている．

3. コロナ禍関連ツイートの取得

本章では，話題分類の対象となるテキストとしてコロナ禍関連ツイートを定義し，取得する．

3.1 ツイートの創作

本稿では，著作権やプライバシーの問題を回避するために，Twitter 上で収集したツイートではなく，アンケート調査によって得られる模擬的なツイートを用いる．

まず，アンケート調査会社である Fastask[14] に登録している 20 代〜50 代のインターネットユーザ 43,221 人（男性 21,490 人，女性 21,731 人）を対象に，Twitter アカウントの有無や Twitter への投稿頻度を尋ね，「たまにつぶやいている」以上の頻度で投稿している Twitter ユーザ 12,326 人（男性 6,364 人，女性 5,962 人）を本稿ではヘビー投稿ユーザとして抽出した．

次に，この 12,326 人のヘビー投稿ユーザからランダムに抽出した 4,346 人（男性 2,202 人，女性 2,144 人）に対し，Fastask 上で「あなたが普段しているようなツイート（140 字以内）を、新型コロナウィルス感染症あるいはコロナ禍の現状に関して今ツイートをするようなイメージで入力してください。」という設問を提示し，一人あたり 1〜2 件のツイートを創作してもらった．その結果，合計で 7,538 件のツイートを得ることができたが，このようにして得たツイートには，無意味な文字列からなるツイートやコロナ禍と全く関係のないツイート，伏せ文字や顔文字，URL，ハッシュタグなどを含むツイート，個人名を含むツイート，文字数制限である 140 字を超えるツイートなどがあった．そこで，データクリーニングを行い，こういったツイートを削除した結果，最終的には 4,615 件のツイートを得ることができた．なお，この 4,615 件のツイートを創作した回答者の異なり数は 2,933 人であった．参考のために，この 2,933 人の性別・年齢構成を表 1 にまとめる．表 1 より，世代間には少しばらつきがあるが，男女間ではほぼ均等になっていることがわかる．

表 1: コロナ禍関連ツイート回答者の性別・年齢

	男性	女性	計
50 代	601	305	906
40 代	547	375	922
30 代	227	505	732
20 代	92	281	373
計	1,467	1,466	2,933

3.2 コロナ禍関連度の評価

前節で得た 4,615 件のツイートを 5 等分し，923 件のツイートからなるデータセットを 5 つ構築した．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（コロナ禍関連度と呼ぶ）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．ここで，コロナ禍関連度の分布を表 2 と表 3 に示し，コロナ禍関連ツイートと認定されたツイートの例を表 4 に示す．なお，表 2 に示したように，コロナ禍関連度が 2.5 以上となるツイートは 2,962 件あったが，fastText によりうまくベクトル化できないツイートが 2 件あったので，この 2 件（「コロナウイルスワクチン」と「コロナキエロ」）は除外した．また，表 3 は表 2 に示した分布の関連度が 2.5 以上の場合の内訳となっている．

表 2: ツイートのコロナ禍関連度の分布

関連度	ツイート数	(%)
2.5〜3.0	2,962	(64.2%)
2.0〜2.5	603	(13.1%)
1.5〜2.0	385	(8.3%)
1.0〜1.5	328	(7.1%)
0.5〜1.0	164	(3.6%)
0.0〜0.5	173	(3.7%)
計	4,615	(100%)

表 3: ツイートのコロナ禍関連度（2.5 以上）の分布

関連度	ツイート数	(%)
3.00	283	(6.1%)
2.89	988	(21.4%)
2.78	938	(20.3%)
2.67	497	(10.8%)
2.56	256	(5.5%)
計	2,962	(64.2%)

4. 話題分類手法

本章では，ツイートを話題分類するための深層学習ベースの手法 [3] について述べる．

手順（1）LDA による話題語の抽出

Latent Dirichlet Allocation（LDA）[4] を用いて前章で得たコロナ禍関連ツイート 2,960 件から 4 つのトピッ

表 4: コロナ禍関連ツイートの例

コロナ禍関連ツイート	関連度
いつまでもマスク生活つらい。数年前の動画なんかを見ると、マスクせずに平気で街の中歩いてるだけでも羨ましく思う。非日常が日常になってしまった今の状況は、いつ終わるのか…	3.00
インフルエンザのように、ワクチンで抑え込める状態にならないと、感染者が増えたり減ったりするのがずっと続くから、一喜一憂しすぎるのは良くないと思う。	3.00
恒例の…とか、例年通り…とかだんだん面倒くさいなって思ってきても、なかなか止め難いけど、コロナを理由に色々止めた！いいこともいっぱいある！	2.89
今緊急事態宣言を解除しても、また増えると思う。今の人数は 12 月の中旬の人数なので、同様のことがまた起こり得る	2.89
コロナはよ去ってくれー。マスクのない生活に戻りたひ	2.78
はやくコロナ収まらないかなあ。「自粛自体」はそんなに苦痛じゃないけど遊びに行けないのは楽しくない。	2.78
国会議員の夜の会食などに対する世間の風当たりが強いが、これが同調圧力になると怖い。	2.67
国民人口あたりの感染確認者数が多くない日本は大した国です。	2.67
マスク外してちゃんとフルメイクして外に出かけたいなあ！	2.56
みんなで集まっておしゃべりしたり、ご飯を食べてワイワイ楽しむことが出来ない。今までは当たり前だった事が出来なくなるのが辛い。	2.56

表 5: LDA の結果（話題語とトピック確率）

（a）トピック 0

早く	0.018
ワクチン	0.016
接種	0.014
なっ	0.013
旅行	0.008

（b）トピック 1

事態	0.050
緊急	0.049
宣言	0.049
感染	0.041
解除	0.032

（c）トピック 2

早く	0.040
人	0.024
マスク	0.023
宣言	0.016
解除	0.016

（d）トピック 3

ワクチン	0.023
者	0.020
感染	0.019
早く	0.015
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: トピックの可視化

クを抽出した（トピック数は正解データ作成時のコストとの兼ね合いから今回は 4 つとした）．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 5 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.「コロナ」を除外したのは，本研究で対象としているツイートがコロナ禍関連ツイートであり，いずれのトピックでも抽出されやすいためである．一方,「する」を除外したのは,「する」がサ変接続の動詞であり，話題語として好ましくないにもかかわらず，様々なサ変名詞に接続して用いられ，結果的に話題語として抽出されやすいためである．

手順（2）ツイートベクトルの生成

前章で得たツイート 2,960 件を fastText[5] を用いてベクトル化した．この 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（単語分散表現）に変換するこ

表 6: 各トピックに分類されたツイートの例（コサイン類似度の最も高かった上位 2 件を抜粋）

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

とができる．今回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，動詞，形容詞，副詞，感動詞を抽出し，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．感動詞を加えたのは,「ありがとう」や「すみません」といった感動詞が重要な役割を果たす場合もあると考えたためである．なお，学習済み日本語モデルには fastText の公式ページ（`https://fasttext.cc/docs/en/crawl-vectors.html`）で公開されている cc.ja.300.vec.gz を用いた．

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．具体的には，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．例えば，あるトピックにおいて抽出された 5 つの話題語のトピック確率を $p_1, p_2, \cdots, p_5$ とし，それぞれの話題語から生成されたベクトルを $\overrightarrow{tv_1}, \overrightarrow{tv_2}, \cdots, \overrightarrow{tv_5}$ とするとき，そのトピックのトピックベクトルは

$$\frac{p_1 \cdot \overrightarrow{tv_1} + p_2 \cdot \overrightarrow{tv_2} + \cdots + p_5 \cdot \overrightarrow{tv_5}}{p_1 + p_2 + \cdots + p_5}$$

となる．

手順（4）メトリックに基づくツイートの話題分類

本研究で採用した深層学習ベースの話題分類手法 [3] では，ベクトルどうしの近さを測るためのメトリックとしてコサイン類似度を用いている．すなわち，それぞれのツイートベクトルにおいて各トピックベクトルとのコサイン類似度を求め，コサイン類似度が最も高かったトピックをそのツイートのトピックとしている．

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高かった上位 2 件のツイートを表 6 に示す．

表 7: ツイートの話題分類作業前に行った教示の内容

ツイートは全部で 2,960 件あります．
それぞれのツイートを読んで，トピック 0，1，2，3 のどれに該当するかを決定します．
各トピックは以下のように 5 つの単語（話題語）によって表されています．

トピック 0　早く，ワクチン，接種，なっ，旅行
トピック 1　事態，緊急，宣言，感染，解除
トピック 2　早く，人，マスク，宣言，解除
トピック 3　ワクチン，者，感染，早く，新型

1 つのツイートに対し，選ぶトピック数はなるべく 1 つにしていただけるとありがたいですが，どうしても 1 つに決められない場合は複数個（2〜4 個）にしてもらっても構いません．

表 8: 正解データ作成に携わった作業者の性別・年齢

	男性	女性	計
40 代	1	1	2
30 代	1	1	2
20 代	0	1	1
計	2	3	5

5. 話題分類の高精度化に関する検討

話題分類の高精度化について検討するためには，話題分類の精度を求めることができなければならない．精度を求めるためには，正解となるデータが必要となる．そこで，まず 5.1 節で各ツイートがどのトピックに属するかを決定し，正解データを作成する．また，前章ではベクトルどうしの近さを測るためのメトリックとしてコサイン類似度を用いる手法を導入したが，本章ではメトリックだけを変え，5.2 節でユークリッド距離を用いる手法を導入し，5.3 節でベースライン手法ともなるトピック確率を用いる手法を導入する．さらに，5.4 節でメトリック（コサイン類似度，ユークリッド距離，トピック確率）の違いが精度に与える影響を調べ，話題分類の高精度化に関して検討する．その結果に基づいて 5.5 節でメトリックの統合を行い，話題分類の高精度化を実現する．

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，クラウドソーシングを利用した．具体的には，20 代〜40 代の男女 5 人の作業者に **3** 章で得た 2,960 件のコロナ禍関連ツイートを 1 件ずつ読んでもらい，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該当すると思うトピックを 1 個以上選ぶという作業を

行ってもらった．このとき，各作業者に示した教示の内容を表 7 に示す．

なお，補足として，トピック 0 の「なっ」は動詞「なる」の活用形であることと先出（左側）の単語ほどそのトピックにおける重要度が高いことを説明した．また，可視化した図に関しては，文字サイズが大きい話題語ほど重要であることと文字サイズ以外の要素（フォントの色や向きなど）はレイアウト上の理由により自動的に決められたものであり，重要度とは関係ないことを説明した．ここで，本作業に携わった 5 人の作業者の性別・年齢構成を表 8 に示す．

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をそのツイートのトピック（正解データ）とした．投票数が同数の場合は，該当するトピックすべてを正解データとした．その結果， 正解データとして 3 つのトピックが選ばれたツイートが 7 件（0.2%），2 つが 293 件（9.9%），1 つが 2,660 件（89.9%）であった．なお，Fleiss' kappa 値 [15] は 0.41 であり，Moderate（中等度の一致）という判定であった．ちなみに，5 人全員の判定が一致したツイートは 680 件（23.0%）と少なめであり，4 人一致が 893 人（30.2%）と 3 人一致が 1,077 人（36.4%）とやや多めで，2 人一致が 310 件（10.5%）となっていることから，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 メトリックとしてユークリッド距離を用いる手法の用意

ベクトルどうしの近さを測るためのメトリックとしてユークリッド距離を用いる手法を作成した．すなわち，それぞれのツイートベクトルにおいて各トピックベクトルとのユークリッド距離を求め，ユークリッド距離が最も小さかったトピックをそのツイートのトピックとすることにした．

5.3 メトリックとしてトピック確率を用いる手法（ベースライン手法）の用意

4 章に示した手順（1）において LDA により話題語を抽出する際，それぞれのツイートに対してもトピック確率がトピック毎に算出される．そこで，各ツイートにおいてトピック確率が最も高かったトピックをそのツイートのトピックとするベースライン手法を作成した．このとき，該当するトピックが複数ある場合は，該当するすべてのトピックをそのツイートのトピックとすることにした．その結果，**3** 章で得たコロナ禍関連ツイート（全 2,960 件）のうち，トピック数が 4 つになったツイートが 15 件（0.5%），2 つになったツイートが 8 件（0.3%）あった．残りの 2,937 件（99.2%）は 1 つのみであった．なお，コサイン類似度とユークリッド距離を用いる手法では，すべてのツイートでトピックは 1 つしか抽出されなかった．

表 9: コサイン類似度を用いたときの混同行列

		コサイン類似度による分類結果				
		0	1	2	3	計
正解ラベル	0	247	0	171	0	418
	1	65	153	919	19	1156
	2	26	12	780	10	828
	3	168	7	570	120	865
	計	506	172	2440	149	3267

表 10: ユークリッド距離を用いたときの混同行列

		ユークリッド距離による分類結果				
		0	1	2	3	計
正解ラベル	0	129	0	288	1	418
	1	3	38	1098	17	1156
	2	4	2	815	7	828
	3	19	1	727	118	865
	計	155	41	2928	143	3267

表 11: トピック確率を用いたときの混同行列

		トピック確率による分類結果				
		0	1	2	3	計
正解ラベル	0	193	25	95	107	420
	1	213	402	376	171	1162
	2	165	75	494	115	849
	3	173	290	176	256	895
	計	744	792	1141	649	3326

5.4 3 つのメトリック間での精度比較

前章で示したコサイン類似度を用いる手法，5.2 節で示したユークリッド距離を用いる手法，5.3 節で示したトピック確率を用いる手法（ベースライン手法）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 値を求め，それぞれのマクロ平均を算出するとともに，正解率を求めた．3 つの混同行列をそれぞれ表 9，表 10，表 11 に示し，各手法の精度を表 12 にまとめる．なお，それぞれの評価指標の定義は以下のとおりとなっている．

$$\text{適合率} = \frac{TP}{TP + FP} \qquad \text{再現率} = \frac{TP}{TP + FN}$$

$$F1\text{ 値} = \frac{2 \cdot \text{適合率} \cdot \text{再現率}}{\text{適合率} + \text{再現率}}$$

表 12: 3 つのメトリック間での精度比較

メトリック	マクロ平均			正解率
	適合率	再現率	F1 値	
コサイン類似度	62.6%	45.1%	37.0%	39.8%
ユークリッド距離	71.6%	36.6%	29.5%	33.7%
トピック確率	39.9%	41.8%	39.3%	40.4%

第 1 位を下線で示す．

$$\text{正解率} = \frac{TP + TN}{TP + FP + TN + FN}$$

但し，本稿では，

TP 〔True Positive〕トピック X と予測したツイートのうち，それが正解であったツイート数

FP 〔False Positive〕トピック X と予測したツイートのうち，それが外れであったツイート数

TN 〔True Negative〕X 以外のトピックを予測したツイートのうち，それが正解であったツイート数

FN 〔False Negative〕X 以外のトピックを予測したツイートのうち，それが外れであったツイート数

と定義する．

表 12 より，全体的な精度（マクロ平均 F1 値）という意味では，コサイン類似度を用いた手法とトピック確率を用いた手法はほぼ同等で，ユークリッド距離を用いた手法は少し低めの値であった．しかしながら，表 9，表 10，表 11 に示した混同行列を見比べてみると，分布傾向がそれなりに異なっていることがわかる．そこで本稿では，この分布傾向の違いを利用して，それぞれの手法を相補的に用いることにより，より高精度な話題分類の実現を目指す．

5.5 メトリックの統合

本節では，3 つのメトリック（コサイン類似度，ユークリッド距離，トピック確率）を統合することで，新たなメトリックを作り出し，それぞれの精度を比較することで，話題分類の高精度化について検討する．

コサイン類似度とトピック確率は，0〜1 の値を取り，近いベクトルどうしでは大きい値を取る．一方，ユークリッド距離は，0〜$+\infty$ の値を取り，近いベクトルどうしでは小さい値をとる．そこで，スケールを合わせるために，ユークリッド距離は逆数にして扱うことにする．但し，ユークリッド距離が 0 でも計算できるように，本稿ではユークリッド距離 +1 の逆数を用いることにする．

まず，この 3 つのメトリックを組み合わせ，以下に示す 4 つのメトリック（順に統合変数 1〜4 と呼ぶ）を作成した．

表 13: トピック確率・コサイン類似度・ユークリッド距離を統合した場合の精度比較

メトリック	マクロ平均			正解率
	適合率	再現率	F1 値	
統合変数 1	41.3%	43.2%	40.0%	41.2%
統合変数 2	66.9%	39.9%	32.7%	35.8%
統合変数 3	40.6%	42.6%	39.8%	41.0%
統合変数 4	40.6%	42.6%	39.7%	40.8%

第 1 位を下線で示す．

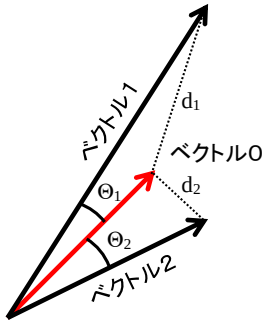

図 2: コサイン類似度とユークリッド距離の関係

$$\text{統合変数 1} = \frac{\text{コサイン類似度} \cdot \text{トピック確率}}{\text{ユークリッド距離} + 1}$$

$$\text{統合変数 2} = \frac{\text{コサイン類似度}}{\text{ユークリッド距離} + 1}$$

$$\text{統合変数 3} = \text{コサイン類似度} \cdot \text{トピック確率}$$

$$\text{統合変数 4} = \frac{\text{トピック確率}}{\text{ユークリッド距離} + 1}$$

それぞれの統合変数をメトリックとして用いる手法を作成し，コロナ禍関連ツイート（全 2,960 件）の話題分類を行った．それぞれの分類結果と正解データから算出される各メトリックを用いた手法の精度を表 13 にまとめる．

表 13 によれば，マクロ平均 F1 値も正解率も統合変数 1 を用いた手法が最も良く，表 12 に示されたベースライン手法（トピック確率を用いる手法）の精度と比べ，マクロ平均 F1 値で 0.7 ポイント，正解率で 0.8 ポイントの上昇が観測された．一方，統合変数間で比べてみると，トピック確率を用いない統合変数 2 では，マクロ平均適合率が高めの値であったが，その他の評価指標は低い値であり，ベースライン手法より悪かった．加えて，統合変数 2 を除く統合変数 1，3，4 では，マクロ平均 F1 値も

正解率もベースライン手法より良くなっており，トピック確率の分類精度への貢献が大きいこととトピック確率とコサイン類似度やユークリッド距離を併用することで分類精度が改善することがわかる．

ここで，コサイン類似度とユークリッド距離のベクトル空間における関係を考えてみる．図 2 に示したように，3 つの（正規化されていない）ベクトルがあり，ベクトル 0 とベクトル 1 のなす角を θ_1，ベクトル 0 とベクトル 2 のなす角を θ_2 とし，ベクトル 0 とベクトル 1 のユークリッド距離を d_1，ベクトル 0 とベクトル 2 のユークリッド距離を d_2 とするとき，$\theta_1 < \theta_2$ および $d_1 > d_2$ という関係が成り立つならば，ベクトル 0 に近いのはベクトル 1 かそれともベクトル 2 かという問題が発生する．この問題はベクトルの大きさが正規化されている場合には考慮する必要がないが，深層学習手法 [5] により生成されるベクトル（分散表現）の大きさは正規化されていない．このようなとき，コサイン類似度とユークリッド距離のどちらがメトリックとして優れているかは一概には言えない．そこで本稿では，この 2 つのメトリックに分類精度への貢献が大きいトピック確率を加えた 3 つのメトリックの線形和という形でも統合変数を作成してみることにする．具体的には，各メトリックに対し係数（α，β，γ）を用意し，次式に示したような統合変数 5 を作成した．

$$\begin{aligned}\text{統合変数 } 5 = {} & \alpha \cdot \text{トピック確率} \\ & + \beta \cdot \text{コサイン類似度} \\ & + \frac{\gamma}{\text{ユークリッド距離} + 1}\end{aligned}$$

この統合変数 5 の係数（α，β，γ）を厳密に求めることはできない．そこで本稿では，試行錯誤することで近似解を求めることにする．具体的には，各係数を 0〜4 の整数とした場合の精度を求めてみた．結果を表 14 にまとめる．但し，紙面の都合により，表 14 にはそれぞれの評価指標において精度が良かった上位 2 件を抜粋して示す．

表 14 によれば，マクロ平均 F1 値も正解率も $\alpha = 1$，$\beta = 4$，$\gamma = 1$ のとき最も良く，表 13 において最も精度の良かった統合変数 1 を用いた手法よりも高い精度になっていることがわかる．表 12 に示したベースライン手法（トピック確率を用いる手法）の精度と比べてみると，マクロ平均 F1 値で 1.7 ポイント，正解率で 2.0 ポイント上昇していることがわかる．しかしながら，今回の試行では，β 値が上限値の 4 であったため，β 値がより大きくなればより高精度になる可能性がある．そこで，β 値を $5 \leq \beta \leq 16$ を満たす整数として統合変数 5 を作成し，それぞれの精度を求めてみた．但し，作業負担を軽減するため，α値とγ値 は $0 \leq \alpha, \gamma \leq 3$ を満たす整数とした．以上の結果を表 15 にまとめる．ここでも，表 14 と同様，それぞれの評価指標において精度が良かった上位 2 件を抜粋して示す．

表 14: 統合変数 5（$0 \leq \alpha, \beta, \gamma \leq 4$）による分類結果の精度（評価指標毎に上位 2 件を抜粋）

メトリック間の	マクロ平均			
$\alpha : \beta : \gamma$	適合率	再現率	F1 値	正解率
0 : 1 : 3	*67.5%*	39.7%	32.8%	35.8%
0 : 1 : 4	68.7%	39.2%	32.3%	35.4%
1 : 4 : 1	44.5%	*45.6%*	41.0%	42.4%
1 : 4 : 2	45.2%	45.8%	*40.8%*	*42.3%*

第 1 位を下線で，第 2 位を斜体で示す．

表 15: 統合変数 5（$0 \leq \alpha, \gamma \leq 4$ かつ $5 \leq \beta \leq 16$）による分類結果の精度（評価指標毎に上位 2 件を抜粋）

メトリック間の	マクロ平均			
$\alpha : \beta : \gamma$	適合率	再現率	F1 値	正解率
0 : 5 : 3	64.5%	43.4%	35.5%	38.2%
0 : 6 : 3	*64.0%*	43.5%	35.8%	38.4%
1 : 7 : 0	50.3%	46.9%	40.0%	42.2%
2 : 9 : 1	45.1%	46.0%	*41.2%*	42.7%
2 : 13 : 0	48.8%	*46.8%*	40.3%	42.4%
3 : 14 : 0	45.1%	46.3%	41.4%	42.9%
3 : 14 : 1	45.3%	46.2%	41.2%	*42.7%*
3 : 14 : 2	45.7%	46.4%	41.2%	*42.7%*

第 1 位を下線で，第 2 位を斜体で示す．

表 15 によれば，マクロ平均 F1 値も正解率も $\alpha = 3$，$\beta = 14$，$\gamma = 0$ のとき最も良く，これまでのどの手法（表 12，表 13 参照）よりも高い精度を達成できていることがわかる．また，表 12 に示したベースライン手法（トピック確率を用いる手法）の精度と比べてみても，マクロ平均 F1 値で 2.1 ポイント，正解率で 2.5 ポイントの上昇となっていることから，メトリックを統合することで，LDA を用いたベースライン手法よりも高精度な手法を実現できることがわかった．

6. まとめ

本稿では，コロナ禍に関連するツイートを対象に，より粒度の細かい話題に分類する深層学習ベースの手法 [3] を示し，その高精度化について検討した．具体的には，アンケート調査により取得した 2,960 件のコロナ禍

関連ツイートを対象に，ツイートとそのツイートから抽出されるトピックをベクトル化し，両者を比較することで，話題分類を行う手法を作成した．このとき，トピック（と話題語）の抽出には Latent Dirichlet Allocation（LDA）[4] を用い，ツイートやトピックのベクトル化には fastText[5] を用いた．また，ベクトルどうしの近さを測るためのメトリックとしては，一般的に用いられているコサイン類似度とユークリッド距離の両方を用いることとし，この 2 つに LDA によって算出されるトピック確率を加えた 3 つのメトリックを統合することで，新たなメトリック（統合変数）を作成した．3 つのメトリックの組み合わせ方を変え，様々な統合変数を作成し，精度を評価した結果，統合変数（$= 3 \cdot$トピック確率 $+ 14 \cdot$コサイン類似度）を用いる手法が最も高精度であり，マクロ平均 F1 値は 41.4%，正解率は 42.9%であった．これは，LDA を用いたベースライン手法（トピック確率を用いる手法）より，マクロ平均 F1 値で 2.1 ポイント，正解率で 2.5 ポイントの上昇となっている．

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，適用範囲のコロナ禍関連ツイート以外への拡張などが挙げられる．コロナ禍関連度の自動判定については，**3** 章で得たコロナ禍関連ツイートとコロナ禍関連度を用いれば，深層学習により実現できる．トピック数の自動決定については，LDA の計算過程で得られる Perplexity や Coherence を用いることで実現できると考えている．コロナ禍関連ツイート以外への拡張については，特定の話題をキーワードとする検索を Twitter 上で行い，取得したツイートに対し実際に試してみることで，その実現可能性を検証したい．また，本稿ではトピック（と話題語）の抽出に LDA を用いたが，対象となるテキストから得られる各単語のベクトルをクラスタリングすることで，トピック（と話題語）を抽出することも可能である [16]．今後の課題としたい．

謝辞

本研究の一部は，JSPS 科研費 20K12085 ならびに福田将治奨学寄付金による研究助成の成果であり，ここに記して感謝の意を表す．

参考文献

[1] 総務省，令和 4 年版情報通信白書，第 2 部第 6 節，https://www.soumu.go.jp/johotsusintokei/whitepaper/r04.html，(2023/4/20 閲覧)．

[2] Twitter，https://twitter.com/，(2023/4/20 閲覧)．

[3] 山下竜也，熊本忠彦，コロナ禍関連ツイートの自動話題分類，第 85 回情処全大，4P-04，2023.

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

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin and Tomas Mikolov, “Enriching Word Vectors with Subword Information,” arXiv:1607.04606v2, 2017.

[6] 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.

[7] 圓谷顯信，上原稔，安達由洋，Sentence-BERT による日本語文の話題分析，FIT2022（第 21 回情報科学技術フォーラム），E-041，2022.

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

[9] David Arthur, and Sergei Vassilvitskii, “k-means++: The Advantages of Careful Seeding,” Proc. of the 18^{th} Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035, 2007.

[10] Nils Reimers, and Iryna Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, arXiv:1908.10084v1, 2019.

[11] Fan Zhang, Wang Gao, Yuan Fang, and Bo Zhang, “Enhancing Short Text Topic Modeling with FastText Embeddings,” Proc. of the International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE2020), Fuzhou, China, pp. 255–259, 2020.

[12] 西田京介，坂野遼平，藤村考，星出高秀，データ圧縮によるツイート話題分類，日本データベース学会論文誌，Vol. 10，No. 1，pp. 1–6，2011.

[13] 王駿キ，佐藤栄一，延原肇，様々なデータ圧縮を用いた多言語に対応する tweets の話題分類法の精度比較，2014 年度人工知能学会全国大会（第 28 回），3M4-03，2014.

[14] Fastask（ファストアスク）モニターサイト，https://monitor.fast-ask.com/，(2023/4/13 閲覧)．

[15] Wikipedia, Fleiss’ kappa, https://en.wikipedia.org/wiki/Fleiss’_kappa, (2023/3/21 閲覧)．

[16] 加藤弘祐，小林弘明，クオリティペーパーを対象とした分散表現に基づくトピック分析－肥満に関する報道の英米比較を事例として－，フードシステム研究，第 28 巻，4 号，pp. 328–333，2022.