社会的話題に関する国内ツイートの データセット作成及び前処理の考察

髙木 裕仁† メンドンサ ドス サントス イスラエル‡ 有次 正義‡

†熊本大学工学部情報電気工学科 〒860-8555 熊本県熊本市中央区黒髪 2 丁目 39 番 1 号
‡熊本大学大学院先端科学研究部 〒860-8555 熊本県熊本市中央区黒髪 2 丁目 39 番 1 号
E-mail: †c8201@st.cs.kumamoto-u.ac.jp, ‡{israel, aritsugi}@ cs.kumamoto-u.ac.jp

あらまし 本研究では多くのツイートデータの中から国内での主要な社会的話題といえる就活と COVID19 のツイートデータに注目する. とりわけ, 日本独自の文化であると考えられる就活に関するツイートデータにおいては, 国内のツイートに対するデータセットの充実を見込めることより，ツイート内容に応じてラベルを付与したデータセットの作成を行なった．また，今回作成した就活に関するデータセットと既存の COVID19 に関するデータセットに対して, 多くの前処理を適用することで, 内容分類を行うタスクにおいて効果的な前処理とは何かを考察する.

キーワード Twitter，ツイート，就活，COVID19，データセット

1. は じ め に

Twitter は日本人が広く利用する SNS の一つであり, 多くの投稿（ツイート）が存在する. ツイートには様々な話題についての内容が含まれており，これらを適切に分析することより何かしらの知見を得られることが期待できる．本論文ではツイート自体に時代の特徴が反映され，多種多様なユースケースに応用が可能である社会的話題の一部である就活と COVID19 の話題を含む日本国内で発信されたツイートを取り扱う.

就活とは，国内では広く定着している就職活動の略語であり，多くの国内の学生が経験することから日本人にとって身近な存在である．これは，諸外国の就職活動と比較すると異なる点が多く，日本独特の制度であるといえる[1]. また, 現在就活を行う上で Twitter は非常に有用なツールとなっており，情報収集において大きな影響力を持っている[2]．これは，学生だけではなく，企業についても同様である．本論文では，就活に関してのツイートのデータセットを作成することにより，国内のツイートにおけるデータセットの充実を期待し，各ツイート内容に応じてラベルを付与したデータセット「就活 2021 日本語 Twitter データセット」の作成を行なった.

また，データセットの作成だけではなく，今回の就活に関するデータセットと既存の COVID19 に関するデータセット[3]に対して，一般的によく知られている前処理や，国内においてのツイートにおいて有用と考えられる前処理を複数適用し， BERT を用いた内容分類を行うタスクにおいて，それぞれの前処理がどの様な影響を与えるのかを考察する.

本論文は以下の様に構成する．第 2 章では，ツイートデータセット並びに本論文で使用している技術の関連研究について述べる．第 3 章では，作成したデータセットの詳細を定義し，第 4 章では，前処理技術に関する考察, 第 5 章では本論文のまとめについて述べる.

2. 関 連 研 究

近年，ツイートをいくつかのクラスに分類し，各クラスに何らかのラベルを付与したツイートデータセットは多く存在する．例えば，SemEval-2017 Task 4 のデータセット[4]ではツイートに 3 つの感情分類を付与している．この様なツイートデータセットは英語のものがほとんどであり，日本語を対象としたものはまだまだ少ないのが現状である．日本語のツイートのデータセットとしては，Twitter 日本語評判分析データセット[5]や COVID-19 日本語 Twitter データセット[3]が存在する．本論文では，日本語ツイートのデータセットの更なる充実を図り，「就活 2021 日本語 Twitter データセット」の作成を行なった.

一方，テキストを扱う機械学習のモデルは Vaswani らの提案した Transformer[6]の考え方をベースにして，目まぐるしい進化を見せている．特に， BERT[7]は人間を超える精度を達成したことで注目を浴びた. Aji ら[8]はツイートデータに対して BERT のいくつかのモデルを利用して内容分類のタスクの比較検討を行なっている．また，それ以降でも続々と T5[9]や GPT-3[10]などの SoTA を更新するモデルが現れている．本論文では，今後より広く一般に使われることが想定される BERT の日本語事前訓練モデルを利用したツイート内

容別に分類を行うタスクに注目する.

テキストデータを扱う際には，モデルだけではなく，モデルの入力の前にどの様にテキストデータに対して前処理を行うことも結果に影響を与える要因の一つである. Symeonidis ら[11]は，２種類のツイートデータセットに多くの前処理を適用して，数種類のアルゴリズムを用いた分類タスクを行って正解率の比較を行った. 結果としては，各処理の精度はツイートデータやアルゴリズムによって異なるものとなった. このことより，どのデータセットやアルゴリズムにおいても汎用的に精度の上昇が期待できるテキストの前処理の存在を断言することは困難といえる.

本論文では，広く知られている前処理や，日本語のツイートの特徴を考慮した前処理などの多くの処理を日本語のツイートデータセットに適用して内容分類タスクを行い，正解率を利用した比較を行うことで各前処理の効果を検討する.

3. デ　ー　タ　セ　ッ　ト

この章では，作成した就活のデータセットについての説明を行う．作成したデータセットは，「就活 2021 日本語 Twitter データセット」である. このデータセットは 2021 年 1 月から 6 月までの半年間の国内で発信されたツイートで，ツイートの中に「就活」という単語を含む 10000 件より構成される．この 10000 件は，日付毎にランダムな時間帯から 500 件を取得し，全てを合わせた半年間（181 日間）における 90500 件のツイートから無作為に選んだものである．今回は，日本の就職活動においての一般的な解禁日は 3 月であり，多くの企業が学生に対して内定を出し終える時期が 6 月であることを考慮し，ツイート活動が頻繁になる上半期の半年間に限定した.

表 1：就活のラベルの対応表

ラベル番号	ラベル名
1	個人の活動状況
2	全体の活動状況
3	企業の利用
4	アドバイス
5	その他

各ツイートには内容別に表 1 のラベルを付与した. 各ラベルの詳細を，そのラベルを設定した理由と共に以下に示す.

ラベル１（個人の活動状況）

特定の個人の活動状況が明記されているツイートを含む．ここでは，ツイートの内容が発信者自身のものでなく，他人の行動に言及したものであっても，個人の活動状況を示していると考えられるツイートを含める．これは，このラベルにツイートから具体的な個人の就活の際の行動を知ることを期待するためである.

ラベル２（全体の活動状況）

一般的な就活生の全体に当てはまると考えられる内容のツイートを含む．このラベルに分類されたツイートは，時期毎や業界毎の就活の特徴を表しているといえる.

ラベル３（企業の利用）

企業が発信しているツイートを含む．近年，求人などの採用活動に Twitter を活用している企業が増加している．また，それに限らず就活生に向けたビジネスの導入として Twitter を利用している場合もある[12]. この様なツイートに共通のラベルを持たせることによって，Twitter 上での企業の利用としてまとめた.

ラベル４（アドバイス）

就活生に対して就職活動についての助言を与えるツイートを含む. 就活生が Twitter を利用する目的として，過去に就活を経験した人が発信している情報を得ることがある．このラベルでは，その様なツイートを表現している.

ラベル５（その他）

上述のどのラベルにも当てはまらないツイートを含む.

ラベル付けは，一つのツイートに，一つのラベルが対応する様に行った．また，一つのツイートの中で複数のラベルの要素を含む場合は，そのツイートの中で最も特徴的と考えられるラベルを対応づけている.

表 2 にツイート例を示す．(a)のツイートは，個人に関する情報を示している. よって, ラベルは 1 とする. (b)のツイートは，前半にアドバイスの要素を含み，後半に個人に関する要素が含まれている．この場合，よりツイートの特徴を表すラベルを適用する. 今回の(b)のツイートは，アドバイスの要素が強く，個人にまつわる情報はアドバイスを根拠づけていると考えられるのでラベルは 4 とする.

表 2：就活に関するツイート例

番号	ツイート	ラベル
a	就活したくない　働きたくない	1
b	ありきたりなことだけれど、就活のうちから結論ファーストで話す癖を身につけるべき。考える習慣が身につく。自分も就活からこれを意識し、実際の仕事でも物凄く役立っている。	4

表 3：COVID のラベル対応表

ラベル番号	ラベル名
1	一般事実
2	個人事実
3	意見・感想
4	関係なし

それぞれのラベルの割合を図 1 に示す．その他以外のラベルはある程度均等に分けられていることがわかる．

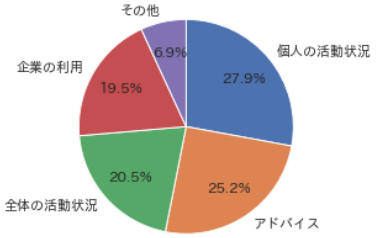

図 1：就活データセットのラベル割合

それぞれのラベルの割合を図 2 に示す．意見・感想のデータが半数以上を占めており，かなりラベルによって偏りがあることがわかる．

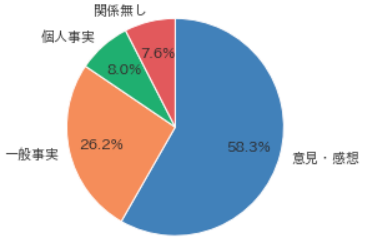

図 2：COVID19 データセットのラベル割合

今回，就活に関するツイートに対してラベル付けを行っている際に，多くのアフィリエイトリンクが見られた．この様なツイートは商用的な利用であるとみなし，3 のラベルを採用している．

4. 前処理に関する実験評価

4.1 データセット

本論文では，今回作成した就活に関するデータセット「就活 2021 日本語 Twitter データセット」と既存の「COVID-19 日本語 Twitter データセット」[3]を利用した．それぞれのデータセットに対して，複数の前処理を適用して内容分類のタスクを行い，正解率に基づいて各前処理の効果を検証する．

COVID-19 日本語 Twitter データセットでラベルが付与されているツイートは 53640 件だが，Twitter の API を利用して，実際にツイートデータを取得できた件数は 37523 件である．また，提供されている元のデータではラベルが 6 つ存在するが，取得できたツイートにおいてはそのうち 2 つのラベルが全体において 1%にも満たないため，本論文では表 3 に示す様に 4 つのラベルを採用した．

4.2 使用モデル

内容分類の際のモデルは，日本語 BERT 訓練済みモデル[13]を用いる．このモデルは日本語版のウィキペディアで事前学習されており，アーキテクチャは Devlin ら[7]の提案した base model と同じである．

4.3 前処理

就活と COVID のそれぞれのデータセットに対して，BERT のモデルに入力する前に表 4 に示す前処理を適用する．前処理の多くは，Symeonidis ら[11]が英語のツイートに対して適用した処理の中から，日本語のツイートにも応用可能なものを採用している．ただし，英文字の除去(10)，文末の「w」や「笑」などの表現を統一する処理(11)，特定の固有名詞の処理(12)の 3 種類は，日本語のツイートの特徴から新たに付け加えた処理である．また，表 4 の前処理名の括弧内の除去はツイートから該当の要素を削除する処理を示し，置換は該当する要素を置き換える処理を示す．

表 4：前処理技術の対応表

番号	前処理名
0	基本（前処理なし）
1	改行（除去）
2	全角空白（除去）
3	ユーザーメンション（除去）
4-a	URL（除去）
4-b	URL（置換）
5-a	ハッシュタグ（#のみ除去）
5-b	ハッシュタグ（一連除去）
6-a	数字（除去）
6-b	数字（置換）
7	絵文字（除去）
8	小文字化
9-a	ストップワード（除去）
9-b	ストップワード（時間系除去）
9-c	ストップワード（漢字系除去）
10	英文字（除去）
11-a	www（w→笑）
11-b	www（笑→w）
12-a	固有名詞（コロナ / 就活）
12-b	固有名詞（ウィルス / インターン）

4.4 学習

表 5 に学習の際に使用したハイパーパラメータを示す．なお，どちらのデータセットにおいても全データを 60%, 20%, 20%の割合でそれぞれ訓練データ，検証データ，テストデータに分割を行う．また，学習の際の正解ラベルは one-hot ベクトルの形で情報を持つ．

表 5：ハイパーパラメーター

項目	ツイート
損失関数	BCEWithLogitsLoss
バッチサイズ	16
学習率	0.00001
最適化関数	Adam
エポック数	10
使用マシン	GPU(Quadro RTX 4000)

4.5 結果

この節では，就活 2021 日本語 Twitter データセットのことを就活， COVID-19 日本語 Twitter データセットのことを COVID と表現する．表 6 にそれぞれのデータセットに対して，前処理を適用し，BERT を用いて内容分類タスクを行った結果を示す．尚，比較を行う際は基本（前処理なし）(0)を基準とし，表 6 の差は(0)の処理との比較した時の差を示す．

表 6：各前処理の適用結果

	COVID		就活	
番号	正解率	差	正解率	差
0	78.97		74.05	
1	78.51	-0.46	73.35	-0.70
2	79.37	0.40	71.30	-2.75
3	77.52	-1.45	74.20	0.15
4-a	76.97	-2.00	67.85	-6.20
4-b	78.72	-0.25	72.70	-1.35
5-a	79.30	0.33	72.80	-1.25
5-b	79.38	0.41	73.90	-0.15
6-a	77.89	-1.09	75.00	0.95
6-b	78.30	-0.66	72.65	-1.40
7	79.26	0.29	73.25	-0.80
8	79.28	0.31	73.00	-1.05
9-a	79.73	0.76	68.20	-5.85
9-b	79.21	0.24	70.50	-3.55
9-c	79.89	0.07	74.25	0.20
10	77.93	-1.04	72.05	-2.00
11-a	79.10	0.13	74.15	0.10
11-b	77.85	-1.10	73.85	-0.30
12-a	78.68	-0.29	71.20	-2.85
12-b	79.88	0.91	70.80	-3.25

改行

改行を取り除く処理はどちらのデータセットにおいても精度が減少した．これは，ラベルによって改行を含むものの割合が異なることによるものと考えられる．この結果より内容分類のタスクでは改行もラベル分類において役に立つ場合があると考えられる．

全角空白

COVID においては空白の除去により精度が上昇し，就活においては精度が大きく下がった.この結果より，就活においては全角空白も分類の際の手がかりの一つとして機能したことが考えられる．

ユーザーメンション

ユーザーメンションの除去においては COVID においては精度が下がり，就活においては精度が向上した．この原因として，COVID に含まれるユーザーメンションが極端に少ないことが考えられる．このことにより，ユーザーメンションそのものが意味を持っていると考えられる．

URL

URL の除去の処理に関しては，両データセット共に最も正解率が下がる結果となった．また，置換の処理

においてはどちらのデータセットも精度自体は下がっているが，除去の処理と比較すると，大幅な現象はみられなかった．

これらの結果は，どちらのデータセットにおいても，異なる日時に発信されているが，同じツイート内容であるものが多く見られたことにより，ラベル予測の際に URL 自体に意味を持ったことと，ラベルにより URL の割合が異なることが処理結果に影響を与えていると考えられる．

ハッシュタグ

ハッシュタグの処理においては，ハッシュタグのみを除去する処理とハッシュタグを含む一連の表現を全て除去する処理の二つを行った．結果としては，一連の表現全てを除去する方がどちらのデータセットに対しても精度が高い結果となった．しかし，ハッシュタグのみの処理と一連の処理のどちらの処理においても，COVID19 では精度が向上し，就活では精度が下がる結果となったことを踏まえると，ハッシュタグの処理の効果はデータセットに依存するといえる．

数字

数字の処理に関しては除去と置換を行なった．比較すると，COVID においては置換の精度が良く，就活においては除去の精度が良い結果となった．1 ツイート当たりの数字の出現頻度は COVID では 1.71 回，就活では 1.77 回となっており，両者に大きな違いはなかったことを踏まえると，数字の持つ意味はデータセットによって異なるといえる．

絵文字

絵文字除去の処理に関しては，COVID の方だけ精度が上がる結果となった．これはデータセットによって絵文字の持つ重要性が異なることを示すと考えられる．

小文字化

小文字化の処理に関しては，COVID の方だけ精度が上がる結果となった．このことから，英語のテキストデータに対してよく用いられる小文字化により表現の次元を下げる処理は，日本語においては必ずしも適切というわけではなく，何かしらの情報を失う可能性があることがわかる．

COVID のデータにおいて精度が上がったのは，出現回数が 2815 回と多かった「covid」という単語の表記の揺れが修正されたことの影響が大きいと考えられる．

ストップワード

基本的なストップワードリストとして SlothLib の日本語ストップワード[13]を用いた．ストップワードリストに含まれる単語をツイートから除去する前処理を適用すると COVID19 の方は精度が向上していることに対して，就活の方は著しく精度が下がる結果となった．

続いて，SlothLib のリストから時間を表す単語（例：「年」「ヶ月」）などの表現を取り除いて，新たなリストとして前処理を行ったところ COVID19 では精度の向上が抑えられ，就活では SlothLib のリストをそのまま適用するより精度の下降が見られない結果となった．

また，SlothLib のリストには多くの漢字を含んだ単語が見受けられることから，リストの中から平仮名のみのストップワードのみを選択し，新たなリストとして前処理を行った．すると，どちらのデータにおいてもわずかながら改善が見られた．これは SlothLib のリストを加工せずに利用した時に，就活において大幅な精度の減少が見られたことと比較すると興味深い結果となった．

以上のことから，前処理にストップワードの除去を行う時には，用いるストップワードリストや適用するデータセットに大きく依存することがわかる．

英文字の除去

この処理では，ツイートから完全に英文字（URL を含む）を取り除いた．こちらの処理においては，完全に URL を取り除いているにも関わらず，URL の除去の処理と比較して，ローマ字全てを取り除いた方がどちらのデータセットにおいても精度が良い結果となった．

www

この処理では，日本語の SNS においてよく用いられる文末の「w」と「笑」に注目し，「w」を「笑」に統一する処理と，「笑」を「w」に統一する処理を実装した．

結果としては，「w」を「笑」に統一する処理においてはどちらのデータセットに対しても多少の精度向上が見られた．一方，「笑」を「w」に統一する処理においては精度が減少した．どちらのデータにおいても「w」を含む割合が多いことを踏まえると，意外な結果となった．これは，URL やユーザーメンションなどで「w」という単語が「笑」という単語よりも，ツイート内で

頻度が高いことが起因していると考えられる.

固有名詞

特定の固有名詞に関して各データセットにおいて,それぞれ二つずつの単語を統一する処理を行なった.処理内容は表 7 のようである.

表 7：固有名詞の処理一覧

データセット	統一処理前	統一処理後
COVID19	covid19	コロナ
	ウィルス	ウイルス
就活	就職活動	就活
	インターンシップ	インターン

それぞれの処理の特徴としては,「COVID19」は専門用語をより日常で普遍的に使われている表現に統一し,「ウィルス」においては統一することにより表記の揺れを修正している.「就職活動」と「インターンシップ」はより使われている省略表現に統一している.

適用結果としては、「COVID19」と「就職活動」と「インターンシップ」においては、精度が下がる結果となった.「COVID19」の結果より,やや専門的な表現と普段よく用いられる表現の統一は好ましくないことがわかる.「就職活動」と「インターンシップ」の結果より,省略表現と省略されていない表現についても同様である.これらより,人間の感覚で同様のものを示すものであっても,安易に固有名詞同士を統一することは注意するべきであることがわかる.

一方,「ウィルス」の処理に関しては他とは異なり精度の向上が見られた.これより「ウィルス」と「ウイルス」といった表記の揺れの問題を解消することは一定の効果が見込まれることが考えられる.

5. ま と め

本論文では,国内で発信された就活に関してのツイートに内容別のラベル付けを行なった「就活 2021 日本語 Twitter データセット」を作成し,作成したデータセットと既存の COVID19 に関するデータセットに対して多くの前処理を適用した.実験結果としては,データセットによって精度の結果が異なるものが多く存在し,前処理の効果はデータセットに大きく依存する傾向が見られた.今後はより多くの異なる話題のデータセットに前処理を適用し,データの特徴などから適した前処理を選択できる手法を考える必要がある.

参 考 文 献

[1] 魚崎典子, ”高等教育機関における外国人留学生のキャリア支援のあり方 : 日本の就職活動の特異性と留学生へのその周知方法をめぐって"多文化社会と留学生交流 : 大阪大学国際教育交流センター研究論集, 18巻, pp. 11-21, 2014

[2] 株式会社 SNS コーチ, ”22 卒就活生の 44.8%が志望企業の Twitter をチェック!約 5 割が企業や社員の Twitter で「志望度が上がった」事実”, PR TIMES, 2021 年 7 月 14 日公開, 最終閲覧日 2022年 1月 5日, https://prtimes.jp/main/html/rd/p/ 000000006.000074113.html

[3] 鈴木優, "鈴木優 : COVID-19 日本語 Twitter データセット", 最終閲覧日 2022 年 1 月 4 日, http://www.db.info.gifu-u.ac.jp/covid-19-twitter-dataset

[4] Sara Rosenthal, Noura Farra, Preslav Nakov, "SemEval-2017 Task 4: Sentiment Analysis in Twitter", In Proceedings of SemEval. Vancouver, Canada, 2017

[5] 鈴木優, "Twitter日本語評判分析データセット",最終閲覧日 2022 年 1 月5 日, http://www.db.info.gifu-u.ac.jp/sentiment_analysis

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, " Attention Is All You Need", arXiv preprint arXiv: 1706.03762, 2017.

[7] Jacob Devlin, Ming-Wei, Chang Kenton, Lee Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", arXiv preprint arXiv:1810.04805, 2018

[8] Alham Fikri Aji, Made Nindyatama Nityasya, Haryo Akbarianto Wibowo, Radityo Eko Prasojo, Tirana Fatyanosa "BERT Goes Brrr: A Venture Towards the Lesser Error in Classifying Medical Self-Reporters on Twitter", Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task, June 2021, Pages 58-64

[9] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", arXiv preprint arXiv:1910.10683, 2019

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, S

am McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei, "Language Models are Few-Shot Learners", arXiv preprint arXiv:2005.14165, 2020

[11] Symeon Symeonidis, Dimitrios Effrosynidis, Avi Arampatzis "comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis", Expert Systems with Applications, Volume 110, 15 November 2018, Pages 298-310

[12] 川畑翔太郎, "コロナ禍で爆伸びするTwitter就活、採用のプロはどう使っているのか？", BUSINESS INSIDER, 2020 年 12 月 7 日公開, 最終閲覧日 2022 年 1 月 10 日, https://www.businessinsider.jp/post-225351

[13] 鈴木正敏, "Pretrained Japanese BERT models", 2019 年 12 月 13 日公開, 最終閲覧日 2022 年 1 月 7 日, https://github.com/cl-tohoku/bert-japanese

[14] Apache Subversion, "slothlibRevision77", 最終閲覧日 2022 年 1 月 7 日, http://svn.sourceforge.jp/svnroot/slothlib/CSharp/Version1/SlothLib/NLP/Filter/StopWord/word/Japanese.txt