

## SNS 世論分析のための一般ドメイン SNS データセットの構築 Open-Domain SNS Dataset for Social Media Public Opinion Analysis

石渡 太智<sup>†\*</sup> 宮崎 太郎<sup>†</sup> 後藤 淳<sup>†</sup>  
Taichi Ishiwatari Taro Miyazaki Jun Goto

### 1. はじめに

放送局では、従来の世論調査で把握できない潜在的な一般世論の動向を明らかにするために、SNS を用いた言論分析を行っている[瀧川 他. 2023]。瀧川らは、図 1 のように、分析対象に関連する SNS 投稿の中で、ネガティブな投稿の割合を推定し、従来の世論調査で得た内閣の不支持率と比較する。ネガティブの割合と内閣不支持率の時間的な変化が異なる場合に、どのような特徴を持つかを分析している。

SNS の世論分析を実現するためには、図 2 のように分析対象に関連する大規模な SNS の投稿を収集し、その投稿が表す感情（ポジティブ、ネガティブなど）を取得する必要がある。しかし、全ての投稿が示す感情ラベルを手で与えるのは困難である。そこで、瀧川らは機械学習モデルを活用し半自動で感情ラベルを取得した。具体的には、約 440 万の投稿から約 1 万の投稿をサンプリングし、人手でポジティブまたはネガティブのラベルを付与し、SNS の投稿とポジティブ / ネガティブラベルの関連性を感情識別器に学習させた。学習した感情識別器に残りの投稿を入力し、その予測値を基に半自動的に約 440 万の投稿の感情ラベルを取得した。

しかしながら、この従来法は、感情識別器が分析対象のドメインに過適合する課題が存在する。例えば、分析対象を当時の”政府”に設定する場合、その分析対象に関連する SNS 投稿で学習された感情識別器は、“コロナ”などの他のドメインに対して汎化性能を保てない。さらに、この従来法を用いて高い識別性能を実現するためには、分析対象ごとに、感情ラベルを新たに人手で付与する必要があり、アノテーションコストが発生する。

そこで、本研究は、分析対象を限定しない一般ドメインの SNS データセットを構築する。さらに、そのデータセットを用いて学習した感情識別器を提案する。一般ドメインデータセットで学習した識別器は、“コロナ”などの分析対象を限定しないため、今後の分析対象ごとの人手アノテーションコストを削減する可能性を持つ。実験結果を通して、提案手法の有効性を確認し、今後のアノテーションコストを削減できる可能性を確認した。

### 2. 一般ドメイン SNS データセット

一般ドメインの SNS データセットを構築する。分析対象を限定せず、幅広いドメインを網羅できれば、今後の新たな分析対象に対して、感情識別器の汎化性能を維持できる可能性がある。

#### 2.1 収集方法

データセットを構築するために、分析対象（検索キーワード）を限定せずに SNS の投稿を収集する。まず 2020 年

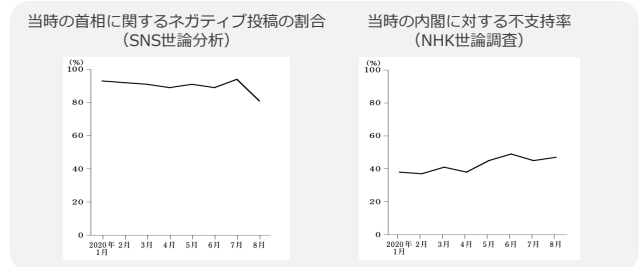


図 1 SNS 世論分析と NHK 世論調査

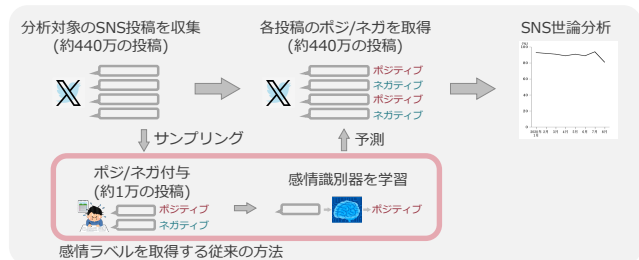


図 2 SNS 世論分析の従来法

2 月 15 日から 2020 年 4 月 14 日までの期間を対象に、X Official Partner である NTT データが提供するなずきのおと分析環境を用いて、X (旧 Twitter) の投稿の 10% サンプルを取得した。次に、感情が現れる投稿を取得するために、絵文字 [Wieslaw et al. 2016] や感情極性辞書 [Ptaszynski et al., 2017] を含む投稿を選択する。さらに、取得した SNS 投稿の文章の品質を向上させるために、広告や不適切な単語を含む投稿を除去する。以上の方法で得られたデータセットを、一般ドメイン SNS データセットとする。

#### 2.2 感情ラベルの付与

各投稿に対して、2 つのアプローチで感情ラベルを付与する。一つは 3 感情ラベルである。“ポジティブ / ニュートラル / ネガティブ” の計 3 つのラベルの中から 1 つを付与する。もう一つは 7 感情ラベルである。“怒り / 嫌悪 / 恐怖 / 喜び / 悲しみ / 驚き / 普通” の計 7 つのラベルの中から 1 つを付与する。投稿の内容が理解できないものを除去するために“判断できない”というラベルも用意する。ポジティブ / ネガティブラベルだけでなく、より詳細な感情ラベルによる SNS 世論分析を行うために、計 7 つのラベルも用意する。また、アノテーションの偏りを防ぐために、同じ投稿を異なる 3 名がアノテーションし多数決でラベルを選択した。計 3 つのラベルの Fleiss' Kappa [Fleiss et al. 1971] は 0.5012、計 7 つのラベルの Fleiss' Kappa は 0.4232 であった。

<sup>†</sup> NHK 放送技術研究所

<sup>‡</sup> 東京工業大学

3 感情	ポジティブ	ネガティブ	ニュートラル	計
コロナ	1,643	3,957	10,499	16,099
一般	50,286	34,038	39,834	124,158

表 1 一般ドメインとコロナデータセット (3 感情)

7 感情	怒り	嫌悪	恐怖	喜び	悲しみ	驚き	普通	計
コロナ	2,198	1,926	2,616	1,490	3,065	533	4,271	16,099
一般	3,957	10,001	3,871	140,073	16,981	5,510	43,765	124,158

表 2 一般ドメインとコロナデータセット (7 感情)

	訓練データ	検証データ	評価データ
コロナ	12,879	1,610	1,610
一般	12,912	1,614	1,614
コロナ+一般	25,791	3,224	3,224
一般 (大)	99,326	12,416	12,416
コロナ+一般 (大)	112,205	14,026	14,026

表 3 各データセットのデータの数

識別器	パラメータの数 (学習パラメータ)
Waseda-RoBERTa	約 0.11B (約 0.11B)
Twitter-BERT	約 0.28B (約 0.28B)
Swallow	約 6.8B (約 0.16B)

表 4 感情識別器のパラメータ数

### 3. 実験

#### 3.1 実験の設定

一般ドメインデータセットを用いて学習した感情識別器の有効性を確認するために、ドメインを限定したデータセットを用いて学習した感情識別器との比較を行う。分析対象を“コロナ”に設定し、本稿で作成した一般ドメインデータセットと別に“コロナ”に関連する投稿を収集し、感情ラベルを付与したコロナデータセットを用意した。一般ドメインデータセットとコロナデータセットの投稿数と各ラベルの出現頻度を表 1 (3 感情) と表 2 (7 感情) に示す。本手法は、一般ドメインデータセットを用いて学習し、コロナデータセットで評価する。これは新規の分析対象に合わせたアノテーションを行わない想定である。一方で、比較手法は、コロナデータセットを用いて学習し、コロナデータセットで評価する。これは分析対象に合わせた新規のアノテーションを行う想定である。

本実験はさらに、データセットの規模の違いを分析する。一般ドメインデータセットのサイズを、コロナデータセットと同程度に揃えたものと、さらに拡張したものを用意する。また、一般ドメインデータセットとコロナデータセットを混ぜて学習した識別器の性能も比較する。表 3 に、コロナデータセット、一般ドメインデータセット、一般ドメインデータセット (大)、コロナ+一般ドメインデータセット、コロナ+一般ドメインデータセット (大) のデータ

の数を示す。感情識別器として、Waseda-RoBERTa<sup>1</sup>、Twitter-BERT<sup>2</sup>、Swallow<sup>3</sup>を利用する。Waseda-RoBERTa と Twitter-BERT はフルファインチューニングを行い、Swallow は LoRA [Hu et al. 2021] チューニングを行う。識別器のパラメータの数を表 4 に示す。

#### 3.2 感情識別器とデータセットの比較

3 つの感情ラベルに対して実験を行った結果を表 5 に、7 つのラベルに対して実験を行った結果を表 6 に示す。識別器の種類と学習に使用したデータセットを行に、評価に用いたデータセットを列に並べる。表の値は 3 回実験を行なった重み付き F1 の平均値を示し、各評価データで最も高い性能をボールド体で示す。

表 5 と表 6 の実験結果から、各評価データでパラメータ数の多い Swallow, Twitter-BERT, Waseda-RoBERTa の順に、認識性能が高くなることを確認した。特に Swallow モデルは他の識別器に比べて飛躍的に性能が向上することを確認した。

次に、データセットの規模がコロナデータセットと同程度の一般ドメインデータセットを検証する。表 5 の 1-2 行目、11-12 行目。表 6 の 1-2 行目、6-7 行目、11-12 行目においてコロナデータセットと一般ドメインデータセットによる結果を比較すると、コロナデータセットの認識性能が高いことを確認した。小規模なデータを用いる場合、評価データとドメインが同じデータで学習する必要があることがわかった。以上の結果から、約 1 万程度のアノテーションを将来的に実施する場合は、分析対象に特化したアノテーションが望ましいことがわかる。

続いて、データセットの規模が大きい場合の結果を比較する。表 5 の 1-4 行目、6-9 行目、11-14 行目の結果から、データセットの規模が大きい場合、評価データのドメインを含まずとも認識性能が向上することを確認した。以上の結果から、約 10 万程度のアノテーションを実施する場合は、分析対象に特化しない一般ドメインデータセット (大) の構築が望まれ、分析対象に特化したアノテーションコストの削減が将来的に期待できる。

最後に、一般ドメインデータセットにコロナデータセットを混ぜて学習した結果を比較する。表 5 の 4-5 行目、9-10 行目、表 6 の 4-5 行目、9-10 行目、14-15 行目の結果から、分析対象をコロナとする訓練データセットを混ぜて学習した方が、性能がさらに改善することを確認した。一般ドメインデータセットだけでなく、評価データと同じドメインの情報が改善に寄与することがわかった。今後は、評価データの投稿内容からそのドメインの情報を予測し、学習時に評価データのドメインに関連する訓練データに重みを付ける手法を検討する。

<sup>1</sup><https://huggingface.co/nlp-waseda/roberta-base-japanese>

<sup>2</sup><https://huggingface.co/Twitter/twhin-bert-base>

<sup>3</sup><https://huggingface.co/tokyotech-llm/Swallow-7b-instruct-hf>

#	3 感情ラベル: ポジティブ, ニュートラル, ネガティブ		評価データ			
	コロナ	一般	コロナ+一般	一般 (大)	コロナ+一般 (大)	
1	コロナ	-	-	-	-	
2	一般	0.4632±0.3040	0.5301±0.1839	-	-	
3	Waseda RoBERTa	コロナ+一般	-	0.7068±0.0185	-	
4	一般 (大)	0.7280±0.0499	-	-	0.7235±0.0152	
5	コロナ+一般 (大)	0.7364±0.0544	-	-	-	
6	コロナ	0.7352±0.0909	-	-	-	
7	訓練 Twitter	一般	0.7587±0.0458	0.7471±0.0272	-	
8	データ BERT	コロナ+一般	-	0.7604±0.0162	-	
9	一般 (大)	0.7618±0.0512	-	-	0.7611±0.0139	
10	コロナ+一般 (大)	0.7693±0.0560	-	-	-	
11	コロナ	0.7985±0.0349	-	-	-	
12	一般	0.7930±0.0251	<b>0.7656±0.0149</b>	-	-	
13	Swallow	コロナ+一般	-	<b>0.7741±0.0235</b>	-	
14	一般 (大)	<b>0.7996±0.0249</b>	-	-	<b>0.7909±0.011</b>	
15	コロナ+一般 (大)	0.7964±0.0219	-	-	-	
					<b>0.7869±0.0102</b>	

表 5 識別器とデータセットの比較 (3 感情ラベル)

#	7 感情ラベル: 怒り, 嫌悪, 恐怖, 喜び, 悲しみ, 驚き, 普通		評価データ			
	コロナ	一般	コロナ+一般	一般 (大)	コロナ+一般 (大)	
1	コロナ	-	-	-	-	
2	一般	0.4883±0.0298	0.5530±0.0462	-	-	
3	Waseda RoBERTa	コロナ+一般	-	0.5499±0.0097	-	
4	一般 (大)	0.4717±0.0474	-	-	0.632±0.0072	
5	コロナ+一般 (大)	0.5123±0.0089	-	-	-	
6	コロナ	0.5461±0.0115	-	-	-	
7	訓練 Twitter	一般	0.4636±0.0508	0.6271±0.0138	-	
8	データ BERT	コロナ+一般	-	0.5971±0.0089	-	
9	一般 (大)	0.5518±0.0388	-	-	0.6703±0.008	
10	コロナ+一般 (大)	0.5539±0.0079	-	-	-	
11	コロナ	0.6097±0.027	-	-	-	
12	一般	0.5647±0.0341	<b>0.6844±0.0197</b>	-	-	
13	Swallow	コロナ+一般	-	<b>0.6431±0.0126</b>	-	
14	一般 (大)	0.5824±0.0442	-	-	<b>0.6997±0.0133</b>	
15	コロナ+一般 (大)	<b>0.6159±0.0156</b>	-	-	-	
					<b>0.6893±0.0088</b>	

表 6 感情識別器とデータセットの比較 (7 感情ラベル)

4. おわりに

放送局では、従来の世論調査で把握できない潜在的な一般世論の動向を明らかにするために、SNS を用いた言論分析を行なっている。SNS 世論分析を実現するためには、分析対象の投稿が示す感情ラベルを取得する必要がある。大規模な投稿の感情ラベルは人手で取得できないので、少量の投稿にラベルを与え学習する識別器を用いた。しかし、識別器は分析対象に過適合するため、他の分析対象に対する汎化性能が低下する課題が存在した。そこで分析対象を限定しない一般ドメイン SNS データセットを構築し、そのデータで学習した感情分析器を提案した。

実験を通して提案手法の有効性を確認し、将来的な Annotations コストを削減できる可能性を示唆した。更なる性能向上のために、今後は分析対象のドメインを投稿内容から判断し、一般ドメイン SNS データセットから特定ドメインの情報を取得し学習に利用する手法を検討する。

参考文献

[1] 瀧川裕貴, 永吉希久子, 呂沢宇, 下窪拓也, 渡辺誓司, & 中村美子. (2023). ソーシャルメディア言論分析の方法 ① 2020 年 1 月から辞任までの安倍首相に対する Twitter 上の投稿分析を事例として. *放送研究と調査*, 73(3), 70-85.

[2] Wolny, W. (2016). Sentiment analysis of Twitter data using emoticons and emoji ideograms. *Studia Ekonomiczne*, 296, 163-171.

[3] Ptaszynski, M., Dybala, P., Rzepka, R., Araki, K., & Masui, F. (2017). ML-Ask: Open source affect analysis software for textual input in Japanese. *Journal of Open Research Software*, 5(1), 16-16.

[4] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 378.

[5] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.