公衆衛生分野におけるオープンソース・インテリジェンスの有効性の検証

安納住子*、木村佳嗣**、杉田暁***
*上智大学、**柳真珠、***中部大学

1. はじめに

日本の気温上昇率は世界平均気温の上昇率（気象庁の解析では 100 年当たり 0.74℃）よりも大きく(1)、また、大都市では都市化による気温の長期的な上昇傾向がみられる(2)。それゆえ、将来の日本においては、温暖化や気温上昇による熱ストレスが増加し、特に高齢者の熱中症リスクが増加することが予想されている。

日本の省庁による熱中症対策においては、救急搬送サーベイランス（搬送者情報：性別、年齢区分、傷病程度、発生場所）および公表、予防に係る普及啓発等が行われている。しかしながら、予防のためには、熱中症の自覚症状をもつ人を早期に検知し、迅速な公衆衛生的対応を可能にするイベントベースサーベイランスが望ましいとされている(3)。近年、ソーシャル・ネットワーキング・サービス(Social Networking Service: SNS）に投稿される発言は、イベントベースサーベイランスの役割を果たすことが期待されている。

本研究は、熱中症に関連する Twitter の投稿文、救急搬送サーベイランスデータをもとに深層学習を用いて、投稿文の事実性分類および公衆衛生学分野におけるオープンソース・インテリジェンス（Open Source Intelligence：OSINT）の可能性について考察することを目的とする。

2. 方法

フレームワークの概要

本研究の目的は、熱中症発生の早期検知を実現するイベントベースサーベイランスを構築することである。早期検知するためのフレームワークを以下に提案する（図 1）。

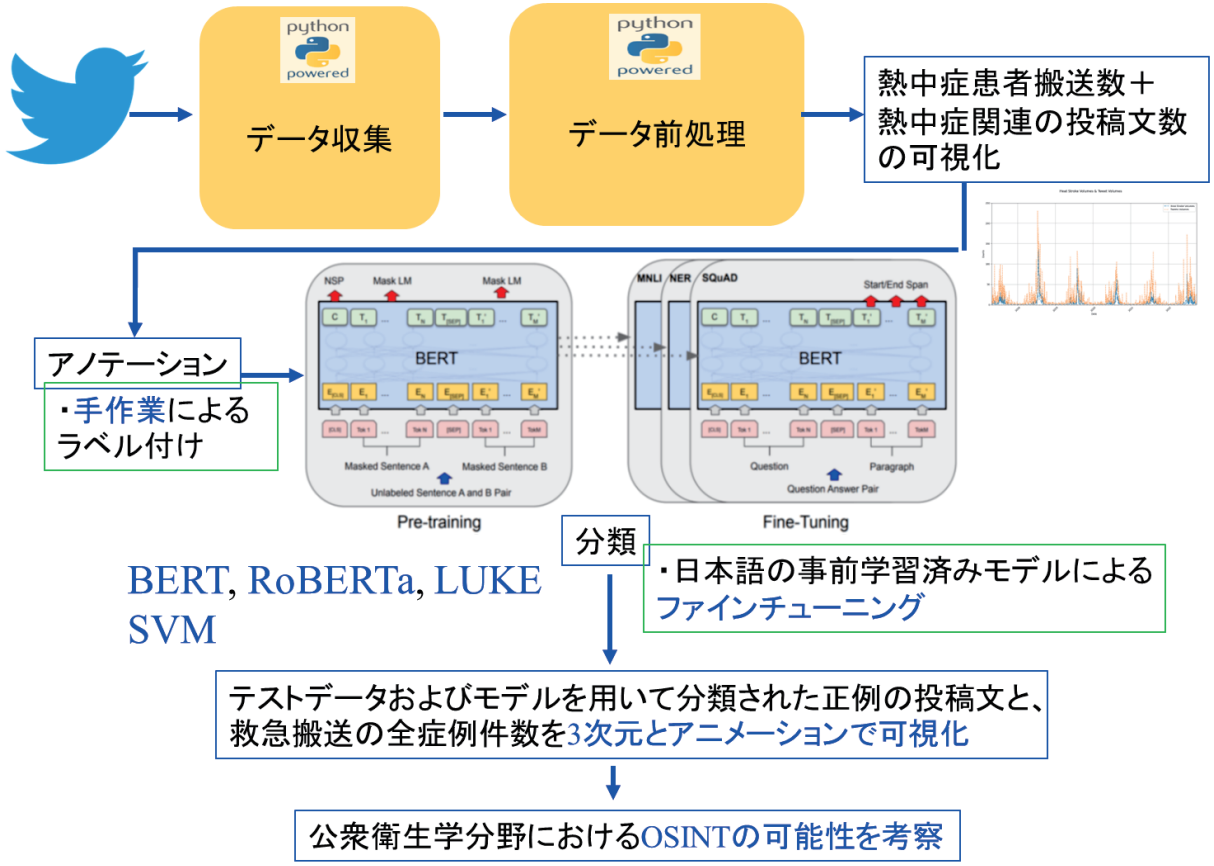

図 1. 研究フレームワークの概要

対象地域

熱中症による救急搬送人員数が、東京都、埼玉県、大阪府に次いで第 4 位(2022 年調べ)(4)の愛知県名古屋市を対象地域とした。

データ収集

本研究に使用する 2 種類のデータを次の通りに収集した。

2017/4/22 から 2022/9/30 までに収集された町丁目レベルの熱中症による救急搬送人員に関するデータを名古屋市消防局から取得した(5)。

Twitter API v2、BEARER_TOKEN を用い、上記の期間に合わせて 2017/4/22 から 2022/9/30 までに投稿された熱中症に関連する投稿文を収集した。

データ前処理

収集した投稿文に対して、クリーニングと正規化の前処理を行うことにより、記号などのノイズを除去した。

時系列の可視化

収集した投稿文が実験に適しているかを確認するため、日毎の熱中症患者搬送数データと日毎の「暑い/あつい」のキーワードを含む投稿文数データを時系列に可視化し、関係性をみた。

アノテーション作業

前処理した投稿文に対してアノテーション作業：熱中症に関連（正例）あるか否か（負例）に分類することで手作業によるラベル付けを行った。

データセットの分割

アノテーションされた正解ラベル付きのデータセットを訓練と検証用に 7.5：2.5 の割合で分割した。

実験

本研究では、分割データ、日本語の事前学習済みモデル: BERT(6)、RoBERTa(7)、LUKE(8)を用いて文書分類の性能を比較した。モデルの評価指標には、正解率、適合率、再現率、F 値を用いた。

時空間上の可視化

精度が最も高かったモデルおよび日毎のテストデータを用い正例として分類された投稿文と、日毎の救急搬送の全症例件をアニメーションで可視化し、両者の関係をみることにより、イベントベースサーベイランスとしての可能性を考察した。

3. 結果

時系列の可視化

日毎の熱中症データ（搬送数）と、日毎の「暑い or あつい」のキーワードを含む投稿文数を時系列に可視化した結果を図 2 に示す。両者の時系列に似た傾向がみられた。

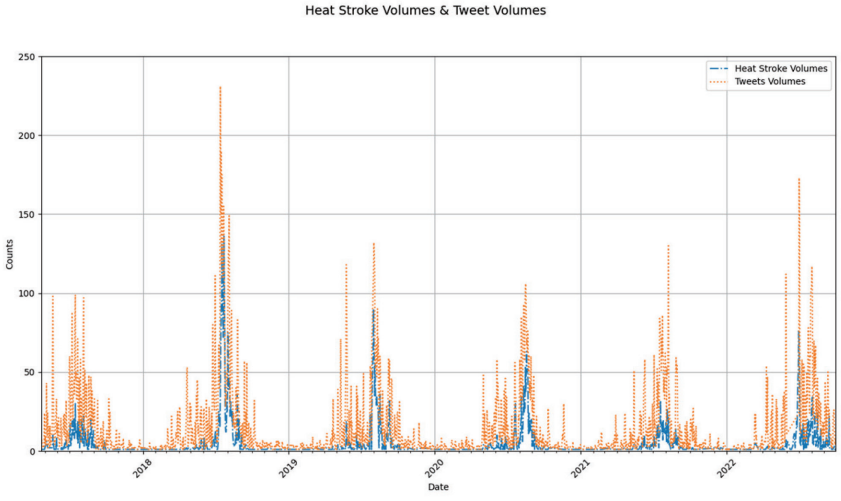

図 2. 熱中症搬送数/日とツイート数/日

アノテーション作業

前処理された投稿文の事例を図 3 に示す。コラム H には前処理前の投稿文、コラム I には前処理後の投稿文が格納されている。記号が完全に除去できておらず、また、丸括弧内に記載した単語や文章も含めてクリーニングされていたことが判明した。これらについてはプログラムを修正し、再度クリーニングを行った。

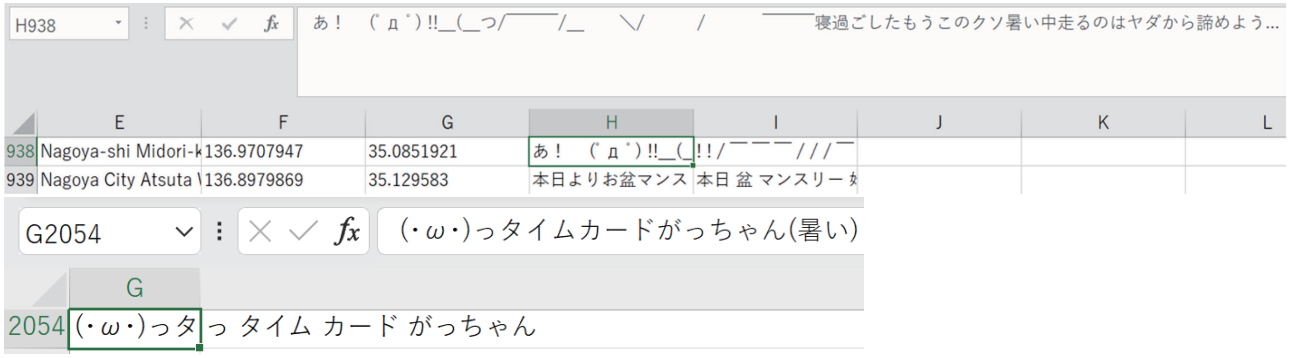

図 3. 前処理前後の投稿文

実験結果

LUKE が最も精度が高く、正解率（Accuracy）が 86.272%、適合率（Precision）が 88.748%、再現率（Recall）が 83.643%、F 値（F-measure）が 86.120%であった。他のモデルにおいてもほとんどの評価指標の値が 80%以上であったことから熱中症に関連（正例）あるか否か（負例）ある程度正確に分類できているといえる。

アニメーションでの可視化

精度が最も高かったモデルおよび日毎のテストデータを用い正例として分類された投稿文と、日毎の救急搬送データをポイントとしてアニメーションで可視化した結果を図 4 に示す。時空間上において、2 種類のポイントが重なり合うことはあまりみられなかったが、互いに近くで分布していた。

図 4. 熱中症搬送数/日と分類結果/日のアニメーションでの可視化：
赤ポイントが投稿文、青ポイントが熱中症発生データ

4. 考察

日毎の熱中症搬送数と「暑い or あつい」のキーワードを含む投稿文数との間に時系列で似た傾向がみられたことから、投稿文の有効性が示唆された。

必要な単語や文章がクリーニングされていたことが判明したため、初年度および今年度はその修正プログラムの作成に時間を費やした。一般の文章とは異なり、投稿文には記号等が多く含まれるため、クリーニングには細心の注意を払う必要があることがわかった。

現時点において最高精度を達成している言語モデル：LUKE は、本研究においても高い精度を達成し、その精度が証明された。

正例として分類された投稿文と、日毎の救急搬送データをポイントとしてアニメーションで可視化し、その時空間上において、重なり合うことはあまりみられなかったが、互いに近くで分布していたことから、投稿文がイベント検知に有効であることが示唆された。

5. まとめ

今後は、アノテーションされた正解ラベル付きのデータセットを訓練、検証、テスト用に分割し、訓練データはファインチューニング の実施に、検証データはモデル作成時のハイパーパラメータの評価に、テストデータは作成したモデルの評価に、それぞれ利用して追加の実験を行う。また、今回使用した日本語の事前学習済みモデル: BERT、RoBERTa、LUKE に加えて、ベースライン：SVM(9)を追加して文書分類の性能を比較する。

上記の実験後、精度が最も高かったモデルおよび日毎のテストデータを用い正例として分類された投稿文と、日毎の救急搬送の全症例件を 3 次元で可視化し、両者の関係をみることにより、イベントベースサーベイランスとしての可能性を考察する。

さらに、高精度で予測するために、分類結果と熱中症データ（搬送数）を用いて熱中症発生リスクを時系列・時空間上で予測するモデルの作成、さらに、熱中症早期警戒システムの構築を目指す。

6. 謝辞

本研究の一部は、中部大学問題複合体を対象とするデジタルアース共同利用・共同研究 IDEAS202309 の助成を受けたものです。また、名古屋市消防局様からは、熱中症救急搬送に関するデータを提供して頂き感謝申し上げます。

参考文献・データ

(1) 気象庁, 世界の年平均気温偏差の経年変化（1891～2022 年),
https://www.data.jma.go.jp/cpdinfo/temp/an_wld.html

(2) Nishat Tasnim Toosty, Aya Hagishima, Ken-Ichi Tanaka, (2021) Heat health risk assessment analysing heatstroke patients in Fukuoka City, Japan. PLoS ONE 16(6): e0253011.https://doi.org/10.1371/journal.pone.0253011

(3) WHO, A guide to establishing event-based surveillance, iSbn 978 9 9061 3 1 3.

(4) 総務省消防庁, 熱中症による救急搬送状況（令和 4 年)「都道府県別救急搬送人員（昨年比)」(グラフ), https://www.fdma.go.jp/disaster/heatstroke/post3.html

(5) 名古屋市消防局.

(6) Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova,(2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

(7) Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

(8) Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, Yuji Matsumoto, (2020) LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention, arXiv:2010.01057.

(9) Corinna Cortes, Vladimir Vapnik,(1995) Support-vector networks. Mach Learn 20, 273-297. https://doi.org/10.1007/BF00994018.