CE-001

テキストデータの極性判定における気象情報の影響の調査
Investigation of the influence of weather information on the polarity determination of text data

山口 匠[1)]　中道 万優子　馬場 隆寛[1)]
Takumi Yamaguchi　Mayuko Nakamichi　Takahiro Baba

1　はじめに

1.1　背景

近年、SNS の利用者数が増加しており、2022 年末の日本での SNS 普及率は 82 %に達した [1]。SNS は情報発信や収集、コミュニケーションのツールとして広く使われており、自身のリアルタイムな情報を共有する場となっている。

一方で、私たちの身近なもののひとつである気候にも注目してみる。近年、地球温暖化の影響から気候変動が起きているといわれており、その影響は測り知れない。そのひとつとしてあげられるのが人間の感情への影響である。気候や気圧などが感情への影響を及ぼし、鬱状態を引き起こすこともある [2]。気候が感情に影響を及ぼすというのは様々な研究で証明されているが、個人差があるというのも事実である。

個人の感情が反映されやすい SNS と、感情に影響を与えている可能性のある気象情報を利用することで感情の起伏をコントロールすることはできれば鬱状態を引き起こす可能性も低くなるだろう。

1.2　目的

以上のような背景から、本研究では X[3] のテキストデータと気象情報を使用して、極性判定を行い、そのデータをもとに感情予測を行う。気象情報を組み合わせることで高精度な感情予測を行うことが明確になれば、日常的に目に入る天気予報や実際の気候によってメンタルへの影響を可視化することが可能になるため、メンタル維持の手助けが期待できるのではないかと考える。また、感情予測の性能をより向上させるためにどの気象情報が感情に影響を及ぼしやすいのかについても調査する。

2　準備

本研究では、X と気象データを使用して実験を行う。X とは、利用者が 140 字以内の文章や写真、動画を通じてコミュニケーションをとる SNS である。以前は「Twitter」という名称だったが、2023 年 7 月に「X」に変更された。令和 4 年度に行われた総務省の調査 [4] によると、全年代での利用率は 45.3 %となっており、最も利用者数の多い 20 代では 78.8 %となっている。幅広い年齢層の人々に利用されており、様々な用途で使用されている。「”いま”を見つけよう」というキャッチコピーの通り、リアルタイムな情報をすぐに収集できる即時性と、情報を爆発的に周知させる拡散性が特徴である。

3　関連研究

これまでも SNS での感情予測や、気象と感情の影響について様々な観点から研究されてきた。本章では、関連する先行研究について述べ、本研究の新規性を明らかにする。

Khalid ら [5] は、畳み込みニューラルネットワークを用いて、ソーシャルグラフネットワークから感情予測をする研究を行った。ここでは、人と人の繋がりや環境、ウェアラブルデバイスから得られるデータ、個人が報告した幸福度やストレス度などの情報を合わせたマルチモーダルデータに基づいて、ユーザの感情状態を予測した。

三井ら [6] は、自然言語処理とナイーブベイズ分類器を用いて SNS に投稿された文章の感情分析を行い、各感情を可視化することのできるライフログシステム「Emote」を開発した。自動的に「怒り・恐れ・喜び・悲しみ・驚き・期待」の 6 種類に分類し、それぞれに色情報を付与してカレンダーと共に表現した。

以上で述べた先行研究では、感情予測を行うことを研究の終着点としている。Khalid らの研究では、周辺環境データの中に天気の情報も含まれているが、それによって感情が左右されるかどうかではなく、あくまで感情予測のためのひとつのデータとして使われている。三井らの研究では、SNS の投稿から感情分析を行っているが、これは自分の感情推移を可視化することが目的であり、どのような要因によって感情が変化するかという記述はない。

また、Baylis ら [7] は Facebook[8] と X を使用し、言語は英語を対象として 35 億件を超える投稿を収集した。そのデータに加え、気候情報「気温・降水量・雨量・湿度」を掛け合わせたデータから気象条件と人間の感情の関係について調査した。Facebook と X でそれぞれ気象が感情に与える影響の傾向の比較も行っている。

以上の先行研究もふまえ、本研究の新規性として、感情予測を研究目的にするのではなく、気候が感情に与える影響について調査する。また、異なる気候情報の種類を取り入れ、英語圏ではなく気候や生活環境が異なる日本での研究を行うことを提示する。

4　提案手法

本研究では、2023 年 2 月 19 日から 2023 年 3 月 12 日の期間、TwitterAPI を使用して、特定の地域のコメントを収集する。同時期の気象データを収集し、X の投稿日時と照らし合わせてマルチモーダルデータを作成する。テキストデータのみと気象データを組み合わせたマルチモーダルデータの両方の極性情報をアノテーションしたデータセットを使用し実験を行う。

4.1　使用したデータについて

2023 年 2 月 19 日から 2023 年 3 月 12 日の期間、TwitterAPI を使用して、位置情報が付与されているもの限定で久留米市役所から半径 20km 以内のコメントを収集した。取集したコメントに対して極性情報のアノテーションを行う。極性判定はポジティブには 0、ネガティブには 1、アノテーションの際にどちらにも当てはまらないものを-1 としその例を表 1 に表記する。

1 日あたり、200 件を収集しており、そのうち極性判定

1）　久留米工業大学大学院 電子情報システム工学専攻

がポジティブなものを 100 件、ネガティブなものを 100 件とそれぞれの極性のデータ数が同じになるようにする。22 日間の収集により、合計 4400 件のデータセットを作成した。

気象データについては、気象庁の過去データをダウンロードできる Web サイト [9] より、X と同期間の収集を行った。気象データは 1 時間ごとに観測されている。以下の表 2 は、使用する気象データの一覧である。

表 1: 極性判定の一例

テキスト	感情	極性
@$\mathbf{S}_K aSi_K u_M aruS$@かすしず さんおはぽむぅ！	うれしい	0
@$\mathbf{m1k}_2$22 ごめんなさい笑 僕が引用したのとスペでお話されてたからだと思います！ みんな優しいので安心してください	悲しい	-1
この企画はアカウント連携が必要なのがネックだったかなぁ	悲しい	1
EverydayTraining 917 日目。 今日はリハビリ行ってスクワットしてきたんで 30 回の 1 セットのみ。昼前からリハビリ行くから夜勤の月曜日は地獄だ...。 https;//t.co/imiatsu9ar	恐れ	1
とにかく西洋は詐欺師ばかり、日本は言うてまだまし、それはホント感じる。	嫌悪	1
..... お前か..... お前がやったのか...	怒り	1

表 2: 使用する気象データ

気象データ	概要
気温 (℃)	地上 1.25~2.0m の大気の湿度
降水量 (mm)	降った雨がどこにも流れ去らずにそのまま溜まった場合の水の深さ
蒸気圧 (hPa)	空気中の水蒸気の分圧
相対湿度 (%)	水蒸気量とそのときの気温における飽和水蒸気との比を百分率で表したもの
日照時間 (時間)	直射日光が地表を照射した時間
風速 (m/s)	10 分間の平均風速

また、X のコメントであるテキストデータの極性判定をアノテーションした一部が表 3 である。

4.2 実験方法

4.2.1 データセットの作成

表 3 のようなデータセットの作成手順は以下の通りである。

表 3: データセットの例

テキスト	感情	極性
お昼はうどんを頂きました。 やっぱ肉ごぼうでしょ〜	うれしい	ポジティブ
2nd のプレミアムがやっぱ１番 最高だったなぁ…… あれこそ神席うれしい	うれしい	ポジティブ
今まで祝日が休みじゃなかったから、祝日見てなくて、 今週祝日があると知って めっちゃ 得した気分になっております	うれしい	ポジティブ
今日のゴルスク楽しくねー 雰囲気悪すぎるや	悲しい	ネガティブ
しつこいアポ取りは嫌われるぜ？ 誰がとか何がとかは言わんけど	嫌悪	ネガティブ
あーイライラすんな。 価値わからんやつは買うな！ 観賞用にした方がましやわ！ ボケェイ！！	怒り	ネガティブ
@kokonananya 特殊詐欺の疑いがあるので、家宅捜索が必要ですね	恐れ	ネガティブ

1）Hugging Face の Transformers ライブラリを使用して、LUKE(LanguageUnderstanding with Knowledge-based Embeddings) モデルを日本語の感情分析に利用する。

2) LUKE モデルを使用し、入力テキストに適した 8 つの感情クラス「”joy(うれしい)”,”sadness(悲しい)”,”anticipation(期待)”,”surprise(驚き)”,”anger(怒り)”,”fear(恐れ)”,”disgust(嫌悪)”,”trust(信頼)”」に分類する。

3) 感情分析の結果を表 4 のように極性判定に変換する。

表 4: 極性判定への変換

ポジティブ	うれしい
ネガティブ	悲しい・怒り・恐れ・嫌悪
ニュートラル (無関係)	期待・信頼

4) 1 日あたり極性判定がポジティブ 100 件、ネガティブ 100 件の計 200 件となるようにデータを収集し 22 日間で 4400 件のデータセットを作成する。このときニュートラルに分類された情報に関しては、データセットに組み込まないこととする。

5) テキストデータのみのデータセットに加えて、気象データの情報を付与したマルチモーダルデータも作成する。

4.2.2 層化 K 分割交差検証を用いた二値分類

1) Jupyter Notebook で python 言語を使用し、機械学習のためのニューラルネットワーク [10] モデルを構築する。

2) テキストデータのみと、気象データを組み合わせた 2 つのプログラムの構築を行う。

3) 機械学習モデルの性能評価手法のひとつである層化 K 分割交差検証 [11] を用いる。理由は以下の通りである。二値分類を行う上でクラスの均衡性が重要になってくるため、層化 K 分割交差検証を用いることで各クラスが均等に分布されモデルの性能評価の信頼性が高くなるからである。また、最終的な評価指標として、正解率 (Accuracy)[12] を使用する。理由として、正解率は「モデルが正しく予測できた割合」として解釈できるため、モデルがどれだけ全体的に良い予測を行えているかを簡潔に表現しているからである。

4) テキストデータのみと気象データを組み合わせたデータセットの性能比較を行う。また、ここでの気象データは「気温・降水量・蒸気圧・相対湿度・日照時間・風速」の全ての要素を使用するものとする。

4.2.3 気象データごとの予測

1)「気温・降水量・蒸気圧・相対湿度・日照時間・風速」の 6 種類の気象情報の組み合わせ計 63 通りの正解率を出力する。

2) 正解率が最も高いものと低いものを比較し、その理由について考察を行う。

5 実験結果と考察

5.1 実験結果

5.1.1 データセットによる正解率の比較

混同行列を表 5 のように表示すると、テキストデータの混同行列は表 6、テキストデータに気象データを組み合わせたデータセットの混同行列は表 7 の通りである。

表 5: 混同行列

予測値＼真値	Positive	Negative
Positine	True Positive (真陽性)	False Positive (偽陽性)
Negative	False Negative (偽陰性)	True Negative (真陰性)

表 6: テキストデータの混同行列

予測値＼真値	Positive	Negative
Positine	1173	1027
Negative	220	1980

表 7: テキストデータ+気象データの混同行列

予測値＼真値	Positive	Negative
Positine	1323	877
Negative	260	1940

テキストデータのみと気象データを組み合わせたデータセットでの正解率は表 8 の通りである。

二つのデータセットを比較すると、気象データを組み合わせたデータセットのほうが約 2.5 %正解率が高いことがわかる。気象が感情に影響を与えており、僅かではあるが感情予測において性能の向上に関与していることが明らかになった。

この場合第 4 章でも表記したように、気象情報として扱うデータは「気温・降水量・蒸気圧・相対湿度・日照時間・風速」の 6 種類である。これらのデータの中で感情予測を向上させる要素や、予測性能を低下させている要素がある可能性が示唆される。

表 8: 感情予測の正解率

使用するデータセット	正解率
テキストデータのみ	**0.71659090**
テキストデータ+気象データ	**0.74159090**

5.1.2 気象情報の組み合わせによる正解率の比較

気象情報 6 種類を組み合わせた計 63 通りの性能比較を行った。

表 9 からは、独立した 6 つの気象情報の正解率、最も正解率の高い「気温・風速」の正解率 0.775、最後に 6 つの項目で比較的正解率の低い組み合わせを表示している。また、最も正解率が低いものは「日照時間」で正解率 0.649 となった。

表 9: 気象データごとの感情予測の正解率

気象データ	正解率
気温	0.710
降水量	0.678
蒸気圧	0.723
相対湿度	0.716
日照時間	**0.649**
風速	0.705
気温 & 風速	**0.775**
降水量 & 相対湿度 & 日照時間	0.708

今回の実験で得られた正解率を散布図で表したものが図 1 である。オレンジ色の三角の点がテキストデータの正解率で、青色の丸の点が気象データを組み合わせたデータセットの正解率である。最も正解率が高いものと低いものでは約 12.6 %の差があるためグラフでもかなり高低差が出ていることがわかる。しかし、それ以外の正解率の値を見てみるとあまり大きな差はなく緩い曲線を描いているように見て取れる。

テキストデータのみの正解率もほぼ中央値であることが分かり、気象データを付与することで感情予測性能に影響を与えることが明らかになった。

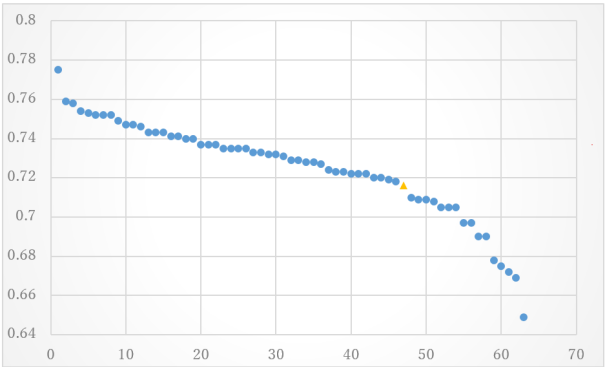

図 1: 正解率の分布

5.2 考察

今回の実験結果から、気象データを加えることで感情予測性能が向上することが明らかになった。また、気象情報の組み合わせによって予測性能に差が生じることから、感情に影響を及ぼしやすいものとそうでない気象情報があるということが分かった。

気象情報の組み合わせによる正解率の比較から、今回性能に最も差が生じた「気温・風速」と「日照時間」について考察を記述する。

「気温・風速」が最も予測精度が高かった理由として、考察は以下のとおりである。今回、2 月 19 日から 3 月 12 日までの期間でテキストデータ及び気象データを収集した。収集時期は暦上春に分類されるが、まだ肌寒い季節でもある。そのため、昼夜での気温差もあり身体的疲労が感情に影響を及ぼすのではないか。また、2 月から 3 月にかけて春一番と呼ばれる強い風が吹くことがあり、普段の生活に少しのストレスが加わる場合や、暖かい風に春の訪れを感じポジティブな感情を抱くことで、感情に動きができると考えられる。

一方「日照時間」の気象情報による予測精度が最も低い理由としては以下のように考える。上記の通り、2 月から 3 月にかけて実験を行ったが、12 月の冬至に近いとはいえ、日照時間が短すぎるということもなく、変化も少ないため感情予測には不要な情報として結果として予測性能が低くなってしまったのではないか。

実験を行う前の個人的見解としては、気圧が感情に与える影響は大きいと感じていた。しかし、今回の実験結果では気圧による予測性能の向上は見られなかった。その理由としてあげられるのは扱っている蒸気圧が関係していると考える。今回気象庁が公開している気象情報のデータを使用したが、現地気圧のデータに関しては気象台が設置してある地域のみ情報があると記載があった。普段、私たちが低気圧・高気圧と目にする情報は「現地気圧」であり「蒸気圧」ではなかったため、正しい気圧の影響については調査することが出来なかったのではないかと考える。

この実験結果から考えられる事柄は以下の通りである。風速の大きい時期には予測性能の向上が見られたが、逆に比較的風速の小さい 7 月から 9 月に実験を行った場合予測性能は低下する可能性が考えられる。また、今回日照時間による予測性能は組み合わせの中で最も低い数値となったが、夏至と冬至では日照時間による影響を受ける可能性があるため、予測性能の向上が考えられる。

また、今回は性能が最も高いものと低いものに注目したが、それ以外の要素に関しても、それぞれの気象情報の影響が強い地域や時期に実験を行うことで予測性能に影響が出る可能性が考えらえる。

6 まとめ

6.1 本論文のまとめ

X の投稿から位置情報が久留米市役所近郊のコメントと同時期の気象データを収集した。収集したコメントは極性判定を行い、テキストデータのみと気象データを組み合わせたマルチモーダルデータの 2 つのデータセットを作成した。

X のコメント (テキストデータ) を使用した感情予測を行う場合、テキストデータのみと比べて気象データを組み合わせたデータセットの予測性能が高いことが明らかになった。

また、気象データの中でも感情に影響をあたえやすいものとそうでないものがあることも明白になり、今回の結果としては正解率の高い組み合わせが「気温・風速」、低いものが「日照時間」となった。

6.2 今後の課題

今回の結論から、情報集時期も感情予測に大きく関係するのではないかと考えられる。そのため、今後の課題としては、1 年を通してデータを収集することで、データ数を増やす目的に加え、感情に影響を及ぼしやすい気象をより明確にしさらに精度を高めていきたい。また、気象情報の値の差異によって感情予測の性能に影響を与えている可能性も考えられるので、それぞれ気象の影響を受けやすい要素を持つ地域に注目して地域ごとのデータ収集を行うことで、さらなる結果を得られると考える。気圧の影響に関しては、気象台が設置してある地域での実験が必要である。

以上の課題をふまえたうえで、当初の目的であるメンタル維持の手助けになるような形のシステムにするためにも、特定の個人のデータを収集することで、気象が感情に影響を与えるのには要素ごとに個人差があるのかというのも調査したい。

参考文献

[1] 出典:ICT 総研. 2022 年度 sns 利用動向に関する調査. https://ictr.co.jp/report/20220517-2.html/. 2024 年 1 月 9 日アクセス.

[2] 西多昌規. 雨の日に気分が落ち込む理由 低気圧とうつ. https://news.yahoo.co.jp/byline/nishidamasaki/20180624-00086868. 2024 年 1 月 9 日アクセス.

[3] X.com. https://twitter.com/. 2024 年 1 月 19 日アクセス.

[4] 総務省情報通信政策研究所. 令和 4 年度情報通信メディアの利用時間と情報行動に関する調査. https://www.soumu.go.jp/main_content/000887660.pdf. 2024 年 1 月 9 日アクセス.

[5] Maryam Khalid and Akane Sano. Exploiting social graph networks for emotion prediction. pp. 1–32, 2021.

[6] 中西勇人 濱川礼三井健史. Sns の投稿を用いた感情記録ライフログシステム〜emote〜. pp. 1–6, 2014.

[7] Yury Kryvasheyeu Haohui Chen6 LorenzoCoviello Esteban Moro Manuel Cebrian James H. Fowler Patrick Baylis, Nick Obradovich. Weather impacts expressed sentiment. pp. 1–18, 2018.

[8] facebook -ログインまたは登録. 2024 年 1 月 19 日アクセス.

[9] 気象庁—過去の気象データ・ダウンロード. https://www.data.jma.go.jp/gmd/risk/obsdl/. 2024 年 1 月 10 日アクセス.

[10] C. M. ビショップ. パターン認識と機械学習 上. 丸善出版, 2012.

[11] Sebastian Raschka 著/Vahid Mirjalili 著/株式会社クイープ 訳/福島 真太朗監訳. [第 3 版]Python 機械学習プログラミング 達人データサイエンティストによる理論と実践. インプレス, 2020.

[12] 門脇大輔, 阪田隆司, 保坂桂佑, 平松雄司. Kaggle で勝つデータ分析の技術. 技術評論社, 2019.