電気通信大学 情報理工学域

令和3年度　卒業論文

Twitter利用者の感情・行動表現の時系列的変動

学籍番号　1710673

氏名　吉田周平

学科　I類 メディア情報学プログラム

指導教員　久野雅樹　教授

提出日　令和3年2月10日（水）

目次

第 1 章

研究背景・目的

1.1　研究背景

SNS の一つである Twitter には、リアルタイムに多くの人々がそれぞれの感情や行動を反映させている。これらのデータは膨大である。

TwitterAPI と Python の形態素解析ライブラリ Mecab を用いることによって、ツイート本文から感情、行動表現を抽出し、時刻やユーザ ID、ツイートタイプ（リプライ、リツートなど)、ツイートソース（bot、アプリケーションなど)、リツート数などの様々な属性情報とともに入手することができる。また、人間は感情を元に意思や行動を決定している。

そこで、時系列変化について変動する Twitter 利用者の感情・行動表現を先に上げた様々な属性情報とともに観察することで、それぞれの時間帯における大衆の心理状態やそれに伴う行動の周期を把握できる。時系列変化に沿った表現の反応を調べるなどして、これらをマーケティング等の分野に役立てることができると考えた。これまでも Twitter を用いて感情や行動の表現の時系列変動を見る研究はあったが、その多くは対象が限定的である。ゆえに、本研究ではより基礎的で普遍的な結果を出すことを目指す。

1.2　研究目的

日本の Twitter 利用者の感情・行動表現に日や週を単位とした周期性があり、これらはユーザの感情・行動の周期性と整合するとを確かめる。

第 2 章

先行研究

2.1　Twitter コンテンツの心理測定指標の日変化

この先行研究 [1] では、2010 年から 14 年までに英国の 54 都市から Twitter に投稿された約 8 億件のツイートと約 70 億の単語を分析し、英国の人々の心理状態が 1 日を通してどのように変化するのかを明らかにした。LIWC[2] と呼ばれる語彙を抽象化してカテゴリ化するためのツールを使用している。

心理的特徴の日周リズムを、24 時間周期の主成分分析により見出した。主成分の 2 つで全体の分散の 85% が説明できることがわかった。第一の要因は、午前 5 時から午前 6 時までをピークとする分析的思考 (あの出来事が起きたのはこういう原因があったはずだなどと分析する思考)、第二の要因は午前 3 時から午前 4 時をピークとする実存的思考 (自分自身の存在意義はなんだ、なぜあれは存在するのだといった思考) であった。

2.2　Twitter を用いた新型コロナ禍における感情変化の分析

この先行研究 [3] では、2019 年に確認された新型コロナウイルス感染症が社会的にどのように話題とされ、どのような影響を人々に与えていたかを明らかにするため、2020/1/17-4/30 の日本の Twitter 上でそれに関連した単語を含むツイートを収集し、人々の関心と感情の変化を分析している。

大きく分けて 2 つの手法を用いて分析を行っている。一つはユーザの偏りの評価、もう一つは感情成分の時間的変化である。どちらの分析からも、3 月の三連休中に存在したと言われる「気の緩み」の存在を示唆する集合現象が観測された。複数の分析から同等の結果が得られたことから、本来定量的には評価できない「気の緩み」の存在がソーシャルメディアから推定できる可能性があることが示唆された。

より長期間の分析、より精度の高い感情分析手法を今後の課題としている。さらに、トイレットペーパーの買い占めが例に挙げられるインフォデミック対策として、情報拡散メカニズムを明らかにすることの重要性を主張している。

第 3 章

コーパスの構築と分析環境の構築

3.1　分析対象

Twitter 社は、常にツイートのうちの 1% を取得できるストリーミング API[4] を公開している。この API を利用し、Python で属性 lang が ja(検出された言語が日本語である) のツイートの以下の属性情報を取得し、json ファイルに蓄積した。

- id
  ツイートの一意の識別子
- text
  ツイートの内容
- created_at
  ツイートが作成された時間 (世界標準時)
- author_id
  ツイートユーザの一意の識別子
- conversation_id
  会話の元のツイートのツイート id(直接返信、返信の返信を含む)
- public_metrics.retweet_count
  リツイート数
- public_metrics.reply_count
  リプライ数
- public_metrics.like_count
  お気に入り数
- public_metrics.quote_count
  引用数
- lang
  検出されたツイートの言語
- source
  ユーザがツイートしたアプリの名前
- referenced_tweets.type
  ツイートのタイプ

3.2　分析期間

2020/12/3 から 2021/2/3 の 2 ヶ月間。収集したツイートの総数は 29251028 ツイートで、一日あたり約 49 万である。

3.3　分析手法

ツイート本文の形態素解析を行い、行動を表す表現として動詞を、感情を表す表現として形容詞を抽出してそれぞれをまとめて蓄積した。形態素解析には、形態素解析システム Mecab[5] を用いた。行った処理は以下の通りである。

1　事前処理
以下のものをツイート本文から除去した。
- @で始まるユーザ名の記述
- #で始まるハッシュタグの記述
- URL の記述

2　形態素解析
形態素解析を行い、ツイート本文を分割した。動詞と形容詞を抽出してリスト化した。

3　時系列データの整形
動詞と形容詞のリストそれぞれについて、以下の処理を行った。

1　時系列データ配列の作成
ツイート登場したすべての単語やツイートそのものについて、その出現数を要素とする以下の次元数のベクトルを作成した。
- 1440（24 × 60）次元の時系列ベクトル（1 分刻みで 1 日ごと）
- 1008 次元の時系列ベクトル（10 分刻みで 1 週間ごと）
- 24 次元の時系列ベクトル（1 時間刻みで 1 日ごと）
- 7 次元の時系列ベクトル（1 日刻みで 1 週間ごと）

2　移動平均化
1440 次元と 1008 次元のベクトルそれぞれについて、前後 5 要素の移動平均をとった。(24,7 次元のベクトルについてはとっていない。)

3　標準化
作成したベクトルそれぞれについて、平均が 0、分散が 1 になるように標準化した。

4　分析
Python のライブラリ tslearn を用いて、時系列データのクラスタリングをおこなった。また、同じく Python のライブラリ sklearn を用いて、主成分分析を行った。

第 4 章

実験 1:ツイート数のクラスタリング

4.1　概要

全体的な数の変化を知るために、ツイートタイプごとのツイート数の日内変動をグラフ化した。また、ツイートアプリごと変化を見ることで、bot など人間以外がつぶやくツイート（定期投稿など）を除外することを試みた。

4.2　ツイートタイプごとのツイート数の日周期変化

4.2.1　手順

1　クラスタリング
ツイートタイプごと（RT、リプライ、引用 RT、それ以外）のツイート数の時系列ベクトル（1440 次元）に対してクラスター数 1-4 としてクラスタリングを行った

2　最適クラスター数の決定
エルボー法を用いた。クラスター数 1-4 において、歪みの大きさを折れ線グラフに出力した。

3　決定したクラスター数でグラフ化
決定したクラスター数において、各クラスターに含まれるツイートの出現数とセントロイドを計算し、1 分刻み（前後 5 分の移動平均)、1 日単位でグラフにした。

4.2.2　結果

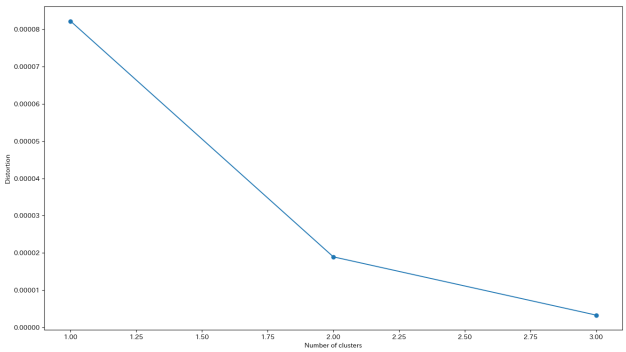

図 4.1. エルボー法による最適クラスター数の決定

クラスター数は 2 が最適であると言える。

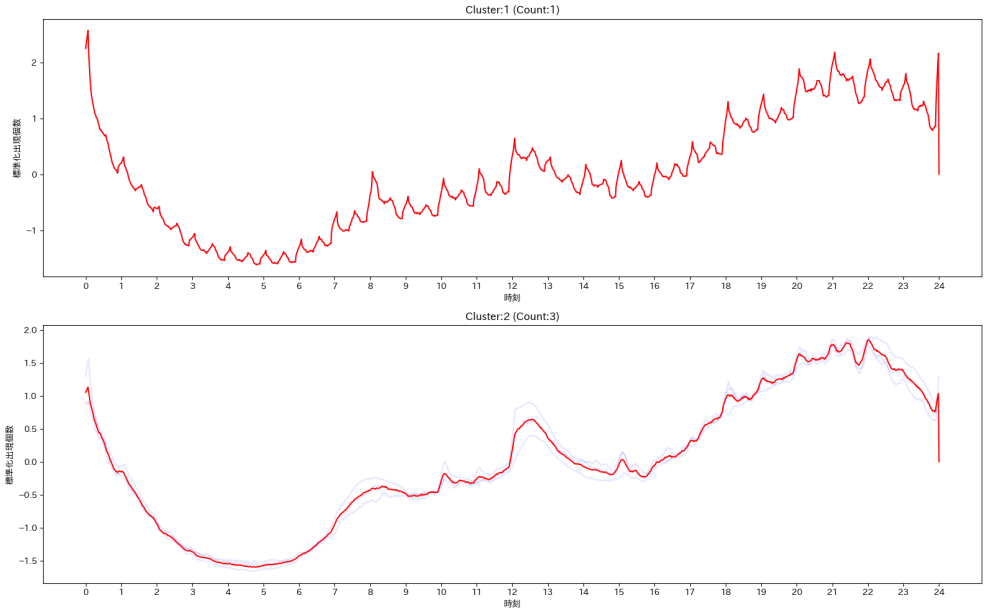

図 4.2. クラスター中心の時系列変化（ツイートタイプ）

4.2.3　考察

クラスター 1 は引用 RT と RT が、クラスター 2 にはリプライと通常ツイートが含まれている。2 クラスターに分かれているが、クラスター 1 がカクついていること以外に大きな差異は見られない。1 が 2 よりもカクついている原因としては、ツイート数が比較的少ないこと、bot やキャンペーン関連のツイートが占める割合が比較的大きいことが考えられる。

4.3　ツイートソース (使用アプリごと) のツイート数の日周期変化

4.3.1　手順

1　分析対象の決定
頻度 10000 以上のツイートソースを抽出した。全 11782 種類から 27 種類に絞られた。

2　クラスタリング
ツイートソースごとのツイート数の時系列ベクトル（1440 次元）に対してクラスター数 1-9 としてクラスタリングを行った。

3　最適クラスター数の決定
エルボー法を用いた。クラスター数 1-9 において、歪みの大きさを折れ線グラフに出力した。

4　決定したクラスター数でグラフ化
決定したクラスター数において、各クラスターに含まれるツイートの出現数とセントロイドを計算し、1 分刻み（前後 5 分の移動平均)、一日単位（1440 分）でグラフにした。

5　クラスターの解釈
各クラスターにおいて、中央値と平均二乗誤差が小さい単語を順に出力し、各クラスターの解釈を行った。

4.3.2　結果

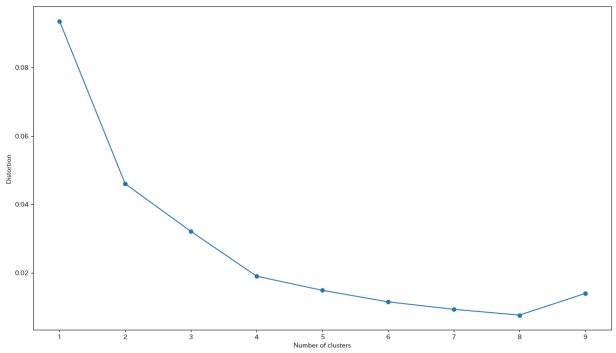

図 4.3. エルボー法による最適クラスター数の決定

クラスター数は 2 とした。

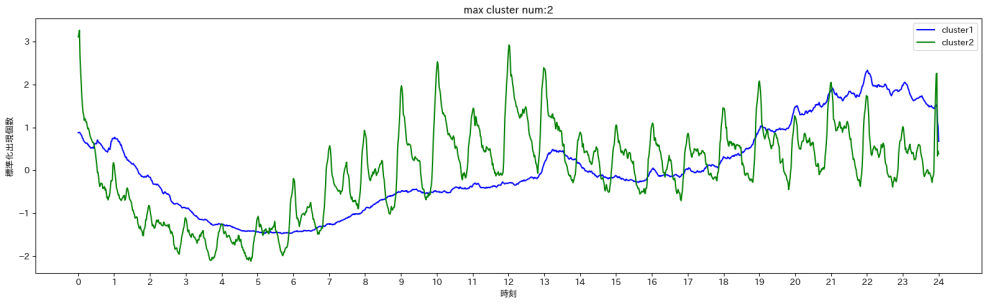

図 4.4. クラスター中心の時系列変化（ツイートソース）

クラスター 1 はツイートタイプで見えた全体の傾向と似ているが、クラスター 2 は一時間単位での増減の周期性が見られる。

表 4.1. クラスター中心からの距離（クラスター 1）

ツイートソース	平均二乗誤差
TweetDeck	0.04353
Twitter for iPad	0.05903
グランブルー ファンタジー	0.07811
Twitter Web App	0.08376
Twitter for iPhone	0.10277
Tween	0.15206
Twitter for Android	0.17341
TwitCasting	0.21281
feather for iOS	0.21634
Echofon	0.21666
Nintendo Switch Share	0.25098
PlayStation®Network	0.28023
Tweetbot for ｉＯＳ	0.28989
TwitPane for Android	0.29144
twitcle plus	0.31379
ツイタマ for Android	0.31398
Peing	0.95827

一般ユーザがにツイートに用いると考えられるアプリケーションが多い。

表 4.2. クラスター中心からの距離（クラスター 2）

ツイートソース	平均二乗誤差
OWNLY Admin	0.43806
Twitter	0.49022
WordPress.com	0.51878
Echoes Act2	0.555
Shuttlerock - Bluebird	0.57808
SocialDog for Twitter	0.61337
Botbird tweets	0.86593
twittbot.net	1.08539
BelugaCampaignSEA	1.1746
今日のツイライフ	1.86464

bot など自動でつぶやけるアプリケーションが含まれている。企業などが用いる定期投稿ができるツールも含まれる。

4.3.3　考察

ツイートをするアプリによってその数の変化に違いが見られた。クラスター 1 には一般的にツイートに用いられるアプリが多く見られ、一日を通してなめらかな変化をしていた。クラスター 2 は bot や定期投稿をするためのツールが多く含まれ、一時間単位で増減の周期が見られた。この結果は 1 時間毎に定期投稿や bot が多くなると考えれば、妥当性が高い。よって以降の分析において、リアルタイムに反映されるユーザの表現でないという点で今回の目的に反するのでクラスター 2 に含まれるツイートソースからのツイートを除外することにした。また、クラスター数を増やしていくと、ゲーム機からのツイートが夜にかけて多くなっていくなど、Twitter が利用者の行動を反映していることの裏付けが得られた。

第 5 章

実験 2:単語頻度のクラスタリング

5.1 概要

動詞と形容詞、それぞれに対して日毎、週毎の時系列変化の波形でクラスタリングと各クラスターの解釈を行った。また、日毎のクラスタリングにおいて、「特異な時系列変化をする単語ほど反応が良い」という仮説のもと、各単語と各クラスターとの距離（平均二乗誤差）と平均お気に入り数、平均リプライ数、平均 RT 数との関係を調べた。

5.2 手順

1 分析対象の決定
4.4.3 におけるクラスター 1 のソースからのツイートに含まれる、頻度 50000 以上の動詞 159 語、頻度 10000 以上の形容詞 131 語を抽出した。

2 クラスタリング
単語の出現数の時系列ベクトル（日毎：1440 次元、週毎：1008 次元）に対してクラスター数 1-9 として K-shape 法を用いてクラスタリングを行った。（1440=24 × 60,1008=7 × 24 × 60 ÷ 10）

3 最適クラスター数の決定
エルボー法を用いた。クラスター数 1-9 において、歪みの大きさを折れ線グラフに出力した。

4 決定したクラスター数でグラフ化
決定したクラスター数において、各クラスターに含まれるツイートの出現数とセントロイドを計算し、グラフ化した。（日毎：1 分刻み/前後 30 分の移動平均、週毎：10 分刻み/前後 30 分の移動平均）

5 クラスターの解釈各クラスターにおいて、中央値と平均二乗誤差が小さい単語を順に出力し、各クラスターの解釈を行った。この論文には小さい方から 20 個を掲載した。

6 各単語の平均反応数と各クラスターとの距離の相関（日毎のみ）各単語の平均お気に入り数、平均リプライ数、平均 RT 数と各クラスターとの相関係数を計算し、散布図を作成した。

5.3　結果

5.3.1　動詞（日変動）

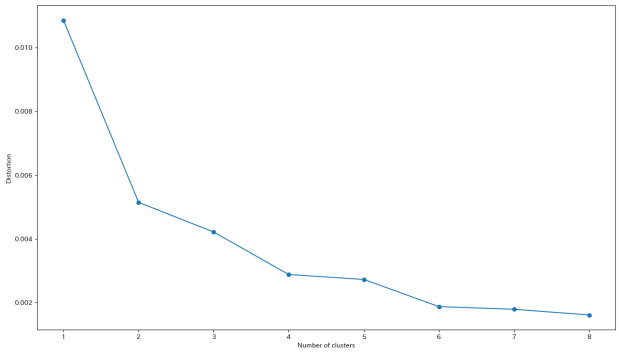

図 5.1. クラスター数の決定

クラスター数は 4 とした。

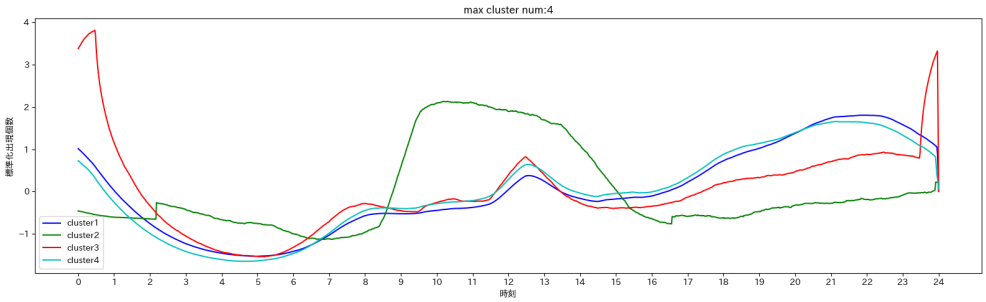

図 5.2. クラスター中心の時系列変化 (動詞、日内変動)

クラスター 1 はクラスター 4 に形が似ている。しかし、クラスター 1 は比較的夜に多く、クラスター 4 は比較的日中に多い。

表 5.1. クラスター中心からの距離（クラスター 1）

単語	平均二乗誤差	分類
やる	0.00205	継続
聞く	0.00224	継続
見る	0.00356	継続
出来る	0.00405	継続
てる	0.00419	継続
分かる	0.00442	瞬間
かける	0.00481	継続
みる	0.00538	継続
く	0.00578	継続
始める	0.00591	瞬間
いう	0.00598	継続
ちる	0.0064	継続
しまう	0.00685	継続
くる	0.00722	継続
やめる	0.00735	瞬間
くれる	0.00797	継続
終わる	0.00955	瞬間
考える	0.0117	継続
読む	0.01588	継続
疲れる	0.03223	瞬間

瞬間動詞が含まれる。瞬間動詞とは、状態の瞬間的な変化を表す動詞である。対して、継続動詞はある程度の動作が継続するものに対して用いる。この論文では、状況によって少しでも時間幅がある動作を示すことがある場合は継続動詞としている。グラフと合わせて、瞬間的な動作表現が比較的夜に多くなることがわかる。

表 5.2. クラスター中心からの距離（クラスター 2）

単語	平均二乗誤差	分類
あける	0.06722	継続
起きる	1.47814	瞬間
答える	2.41437	継続

キャンペーンや起床、挨拶メッセージに関わる単語が含まれる。

表 5.3. クラスター中心からの距離（クラスター 3）

単語	平均二乗誤差	分類
置く	0.04234	継続
届く	0.06577	瞬間
当たる	0.17622	瞬間
寝る	0.52132	継続

キャンペーンや就寝に関わる単語が含まれる。

表 5.4. クラスター中心からの距離（クラスター 4）

単語	平均二乗誤差	分類
いる	0.00169	状態
する	0.00262	継続
れる	0.00377	継続
増える	0.00421	継続
ある	0.00426	状態
乗る	0.00434	継続
持つ	0.005	継続
出る	0.00628	継続
使う	0.00637	継続
できる	0.00673	継続
動く	0.00689	継続
かかる	0.0074	継続
下さる	0.00946	継続
なる	0.01019	継続
いく	0.01451	継続
くださる	0.0165	継続
書く	0.01768	継続
受ける	0.01995	継続
いたす	0.07142	継続
頑張る	0.19334	継続

クラスター 1 が瞬間動詞を含むのに対して、こちらは状態動詞を含む。状態動詞とは、ある状態にあることを示す動詞である。グラフと合わせて、状態を表す動作表現が比較的日中に多いことがわかる。

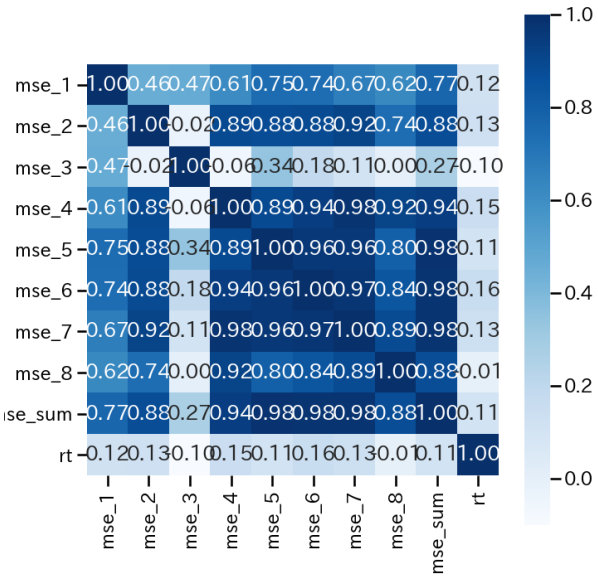

図 5.3. クラスター中心からの距離と平均 RT 数の相関係数

単語毎のクラスターとの距離と平均 RT 数に相関はなかった。

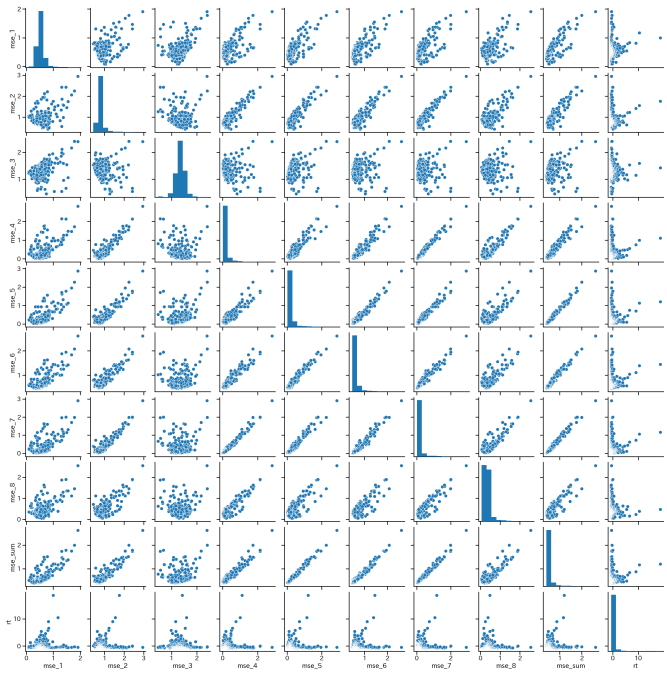

図 5.4. クラスター中心からの距離と平均 RT 数の散布図

相関がないことがわかる。お気に入り数、リプライ数、引用数に関しても同様に相関は認めらなかった。

5.3.2　動詞（週変動）

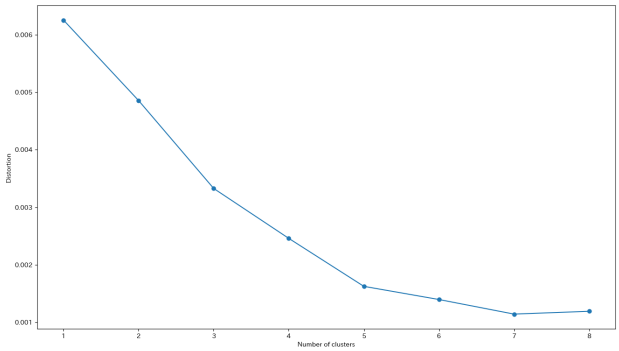

図 5.5. クラスター数の決定

こちらからは 5 も良いように見えるが、日変動との対応を観察するためにクラスター数は 4 とした。

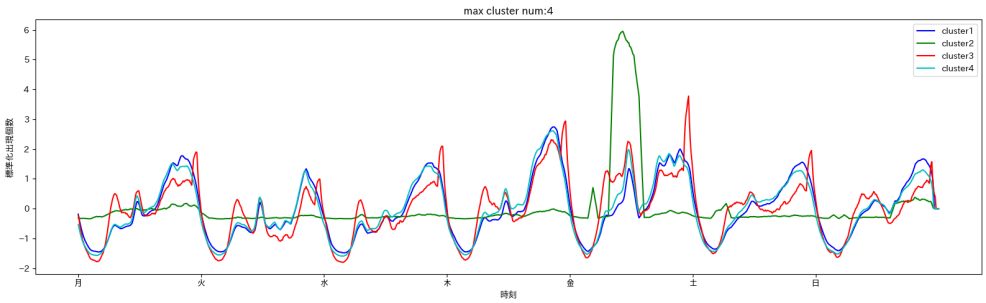

図 5.6. クラスター中心の時系列変化 (動詞、週内変動)

クラスター 2 が金曜に突出して多い。クラスターそれぞれが日変動の変化と対応しているが、クラスター 2 のみ、元旦の挨拶ツイートの影響を強く受けている。

表 5.5. クラスター中心からの距離（クラスター 1）

単語	平均二乗誤差	分類
みる	0.14585	継続
始める	0.16241	瞬間
出来る	0.17141	継続
く	0.1724	継続
てる	0.17494	継続
かける	0.18117	継続
やる	0.18293	継続
聞く	0.18329	継続
思う	0.18366	継続
いう	0.18774	継続
疲れる	0.18933	継続
やめる	0.19114	瞬間
分かる	0.19311	瞬間
ちる	0.20556	継続
見る	0.20902	継続
終わる	0.21317	瞬間
くれる	0.21492	継続
しまう	0.21678	継続
答える	0.48109	継続
寝る	1.6172	継続

瞬間動詞が含まれる。日内変動のクラスター 1 に対応している。

表 5.6. クラスター中心からの距離（クラスター 2）

単語	平均二乗誤差
あける	0.2252

時系列変動において、金曜のみに突出して多かったことから、「あけましておめでとう」によるものだと考えられる。

表 5.7. クラスター中心からの距離（クラスター 3）

単語	平均二乗誤差	分類
頑張る	0.24755	継続
過ごす	0.26401	継続
がんばる	0.29662	継続
当たる	0.73403	瞬間
置く	0.77284	瞬間
起きる	0.82186	瞬間
届く	1.10398	瞬間

瞬間動詞が多く含まれる。日内変動のクラスター 2、3 に対応し、起床やキャンペーンによるものが含まれる。

表 5.8. クラスター中心からの距離（クラスター 4）

単語	平均二乗誤差	分類
増える	0.19466	継続
いる	0.20617	状態
する	0.20905	継続
乗る	0.21178	継続
ある	0.21374	状態
持つ	0.22627	継続
れる	0.22632	継続
使う	0.23034	継続
受ける	0.2304	継続
できる	0.2307	継続
くださる	0.23263	継続
出る	0.24484	継続
かかる	0.24645	継続
いたす	0.2474	継続
くる	0.25095	継続
いく	0.25737	継続
なる	0.25954	継続
書く	0.29239	継続
考える	0.31097	継続
読む	0.3696	継続

日内変動のクラスター 4 に対応し、同じく状態動詞を含む。

5.3.3 形容詞（日内変動）

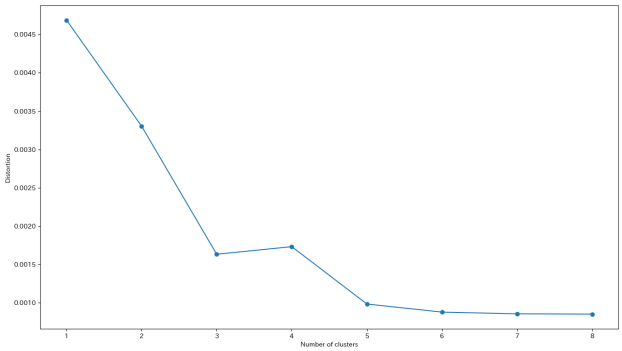

図 5.7. クラスター数の決定

クラスター数は 3 とした。

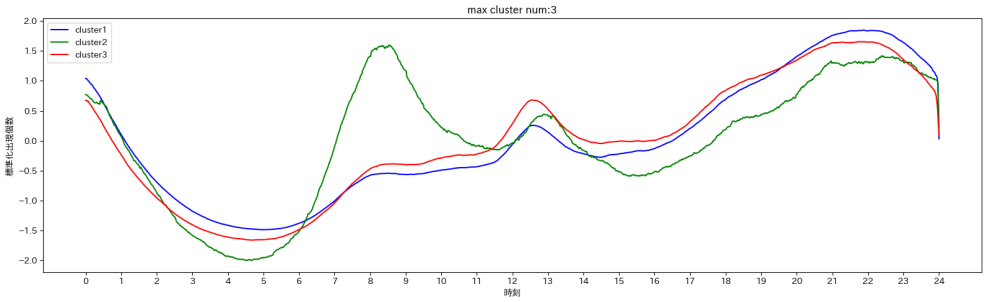

図 5.8. クラスター中心の時系列変化 (形容詞、日内変動)

クラスター 1 とクラスター 3 の形は似ているが、クラスター 1 は比較的夜に多く、クラスター 3 は比較的朝から夕方にかけて多い。クラスター 2 は朝に非常に多くなっている。

表 5.9. クラスター中心からの距離（クラスター 1）

単語	平均二乗誤差
すごい	0.00225
いい	0.00386
悲しい	0.00629
仲良い	0.00645
よい	0.0067
かわいい	0.00772
やばい	0.00851
優しい	0.00909
難しい	0.01
恥ずかしい	0.01003
ありがたい	0.0107
良い	0.01318
細かい	0.01352
うまい	0.0178
こわい	0.02088
素晴らしい	0.02166
欲しい	0.03511
楽しい	0.06224
おもろい	0.06445
たのしい	0.12126

数値化や比較の難しい、主観的な形容詞が多くを占める。グラフと合わせて、主観的な表現が夜に多くなることがわかる。

表 5.10. クラスター中心からの距離（クラスター 2）

単語	平均二乗誤差
暖かい	0.12319
ねむい	0.16763
寒い	0.25176
眠い	0.28412
さむい	0.30496

温度感覚や眠気に由来する形容詞で構成される。グラフより、これらが朝に非常に多くなることが分かる。

表 5.11. クラスター中心からの距離（クラスター 3）

単語	平均二乗誤差
高い	0.00221
多い	0.0027
大きい	0.00321
少ない	0.00436
新しい	0.00442
無い	0.00449
低い	0.00697
近い	0.00985
厳しい	0.00987
ない	0.01016
宜しい	0.0107
薄い	0.01163
悪い	0.0121
遠い	0.01283
危ない	0.01667
若い	0.01731
安い	0.01921
偉い	0.02061
美味しい	0.03646
詳しい	0.10572

数値化や比較の簡単な、客観的な形容詞が多くを占める。グラフより、客観的な表現が比較的昼に多くなることがわかる。

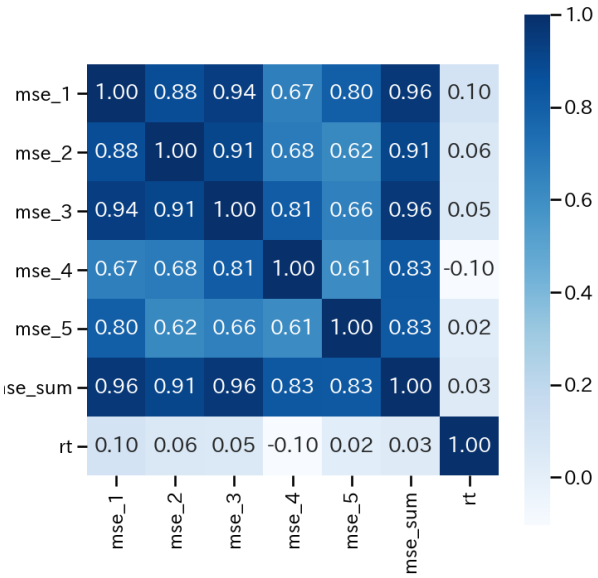

図 5.9. クラスター中心からの距離と平均 RT 数の相関係数

動詞と同様、こちらにも相関は見られなかった。

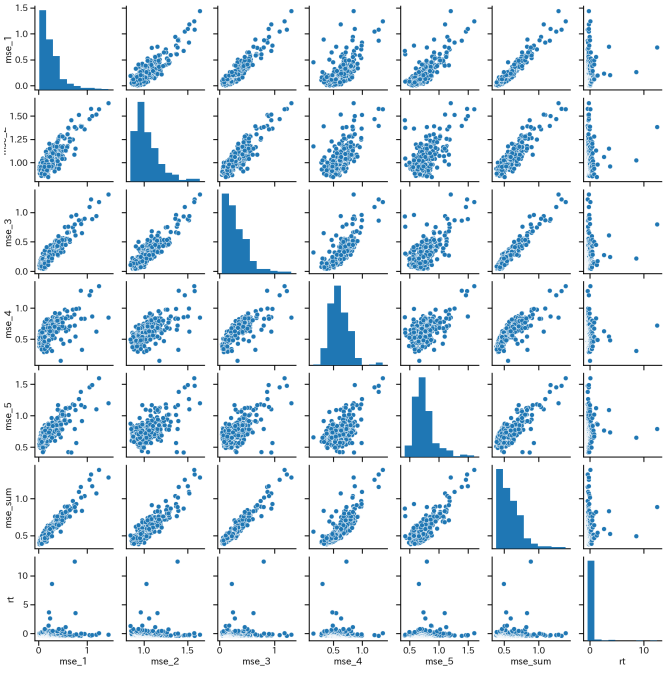

図 5.10. クラスター中心からの距離と平均 RT 数の散布図

同様に、お気に入り、リプライ、引用 RT との相関も見られなかった。

5.3.4 形容詞（週変動）

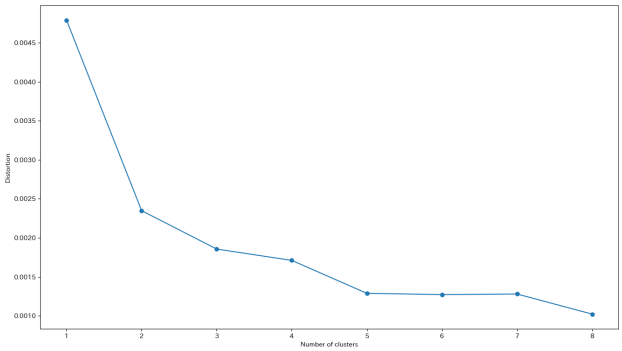

図 5.11. クラスター数の決定

クラスター数は 3 とした。

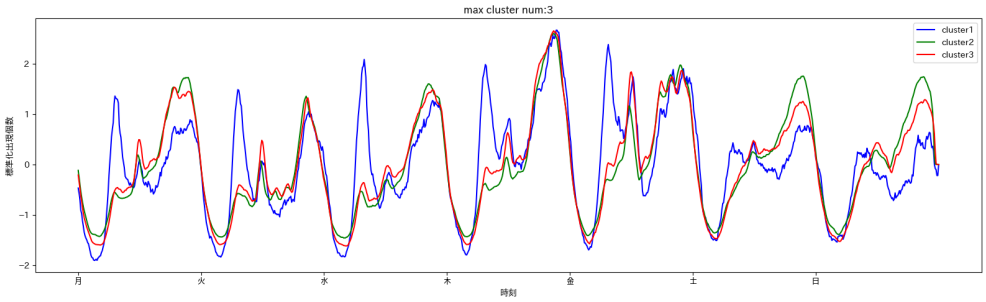

図 5.12. クラスター中心の時系列変化 (形容詞、週内変動)

日内変動と平日の変動の仕方と構成単語が対応している。クラスター 1 は平日の朝に多いが、土日の朝は少なく、クラスター 2 や 3 との大小関係が逆転している。休日に仕事がない人が朝に眠気や気温についてつぶやくことが減ることが原因であると考えられる。

表 5.12. クラスター中心からの距離（クラスター 1）

単語	平均二乗誤差
暖かい	0.1311
寒い	0.13333
さむい	0.25724
ねむい	0.43306
眠い	0.60651

日内変動のクラスター 2 に対応している。

表 5.13. クラスター中心からの距離（クラスター 2）

単語	平均二乗誤差
欲しい	0.10518
いい	0.14962
うまい	0.15762
よい	0.16256
ありがたい	0.16495
良い	0.16652
すごい	0.17953
難しい	0.18087
やばい	0.18514
仲良い	0.18694
かわいい	0.19316
素晴らしい	0.19731
優しい	0.22184
恥ずかしい	0.23916
悲しい	0.24369
こわい	0.25842
細かい	0.27085
楽しい	0.39331
おもろい	0.51241
たのしい	0.56695

日内変動のクラスター 1 に対応している。

表 5.14. クラスター中心からの距離（クラスター 3）

単語	平均二乗誤差
安い	0.09165
少ない	0.09291
高い	0.09891
新しい	0.10125
大きい	0.10563
多い	0.10689
厳しい	0.11192
無い	0.12637
若い	0.12843
ない	0.13415
低い	0.1351
宜しい	0.13918
悪い	0.14193
早い	0.15145
近い	0.15698
痛い	0.17316
遠い	0.17806
薄い	0.18806
偉い	0.19361
危ない	0.21106

日内変動のクラスター 3 に対応している。

5.4　考察

同じ意味の表記が違う単語や対義語が同じクラスターの平均二乗誤差が近い位置に属していることが多いことから、意味と時系列変化に関係は深く存在することが確認できると考えられる。

1 週間でクラスタリングするとたいていのクラスターで 1 日毎に周期ができることから、1 日ごとに単語の時系列変化とその単語の意味や特徴との関係を観察することは意味があることだと考えられる。

週の変動から、どのクラスターも一日毎に強い周期があることがわかる。また、日変動が週変動と強く対応している。

温度感覚や眠気に関する形容詞は、平日と休日で異なる変化をする。

形容詞には朝から夜にかけて分析から解釈という表現の変化が現れている。

動詞には朝から夜にかけて継続的か瞬間的（変化的）という表現の変化が現れている。

また、特異な変化をする単語は希少であることが考えられるので、これらに対するユーザの反応を調べるためにはより多くのツイートと単語が必要である。

第 6 章

実験 3:単語頻度の主成分分析

6.1　概要

動詞と形容詞、それぞれに対して、1 時間ごとの出現数と 1 日ごとの出現数を説明変数として主成分分析を行った。

6.2　手順

1　分析対象の決定
頻度 5000 以上の動詞 1721 語、頻度 1000 以上の形容詞 287 語を分析対象とした。

2　主成分分析
単語の出現数の時系列ベクトル（日毎：24 次元、週毎：7 次元）に対して主成分分析を行った。

3　主成分選択
累積寄与率を計算し、第 2 主成分までを選択した。

4　主成分得点の計算
単語それぞれに対して第 1 主成分得点と第 2 主成分得点を散布図で図示した。主成分得点が高い 10 単語について、対応する時間幅でグラフ化した。

5　主成分の解釈
第 1、第 2 主成分それぞれに対して観測変数の寄与度を散布図で表し、主成分の解釈を行った。

6.3　結果

6.3.1　動詞（日変動）

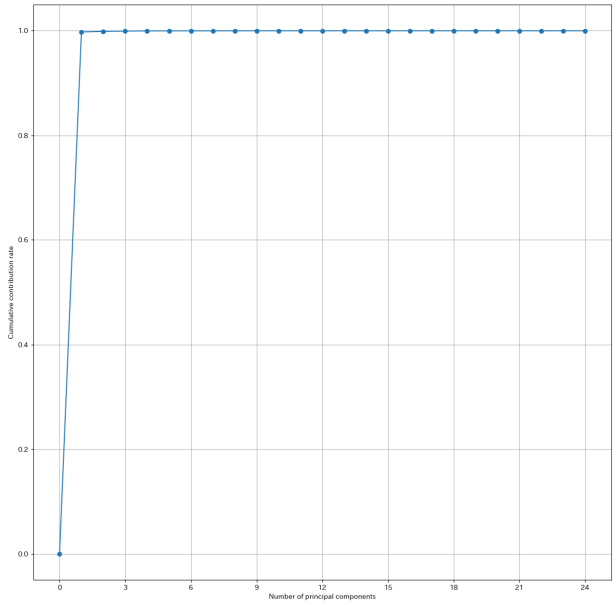

図 6.1. 累積寄与率

ほぼ第一主成分のみで説明できてしまっている。

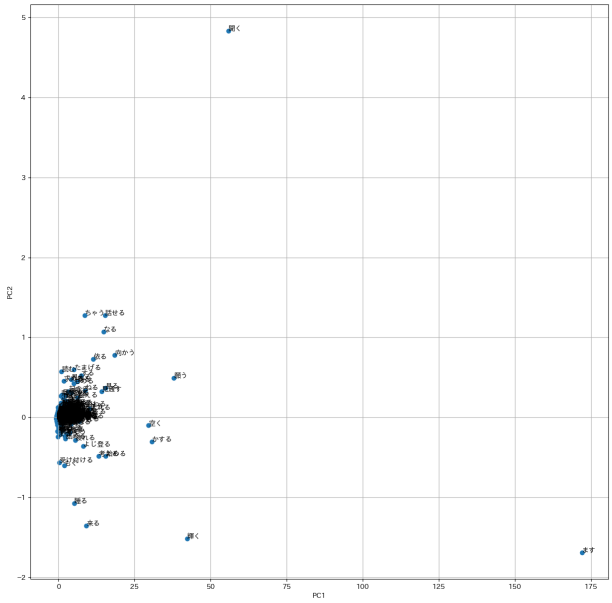

図 6.2. 主成分得点

原点付近に集中している。解釈が難しい。

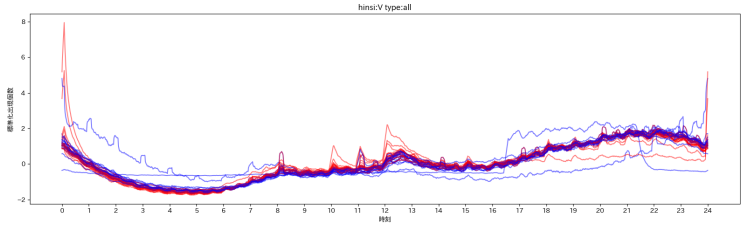

図 6.3. 主成分得点上位 10 単語の時系列変化

赤は PC1 が高い単語、青が PC2 が高い単語である。

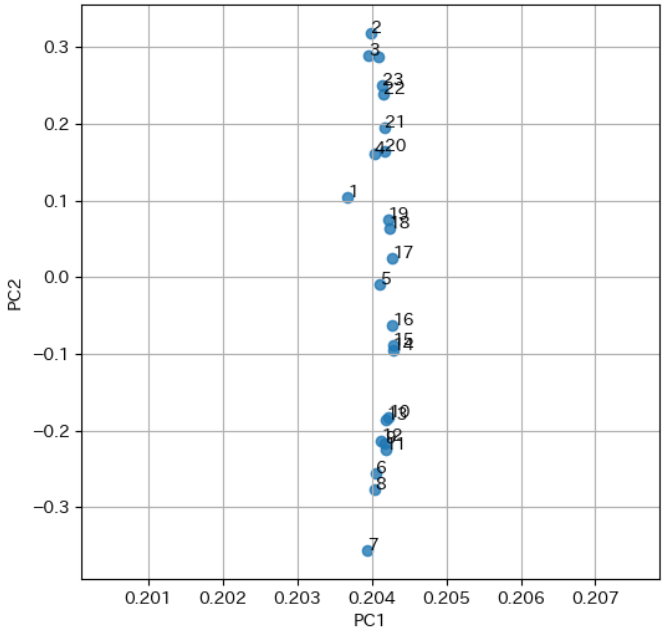

図 6.4. 主成分に対する観測変数の寄与度

PC1 は全体から満遍なく寄与されている。対して PC2 は早朝と深夜から寄与されている。

6.3.2 動詞（週変動）

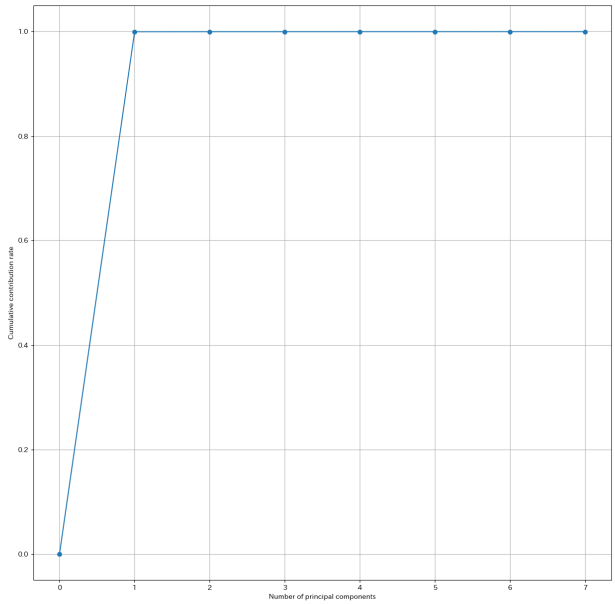

図 6.5. 累積寄与率

ほぼ第一主成分のみで説明できてしまっている。

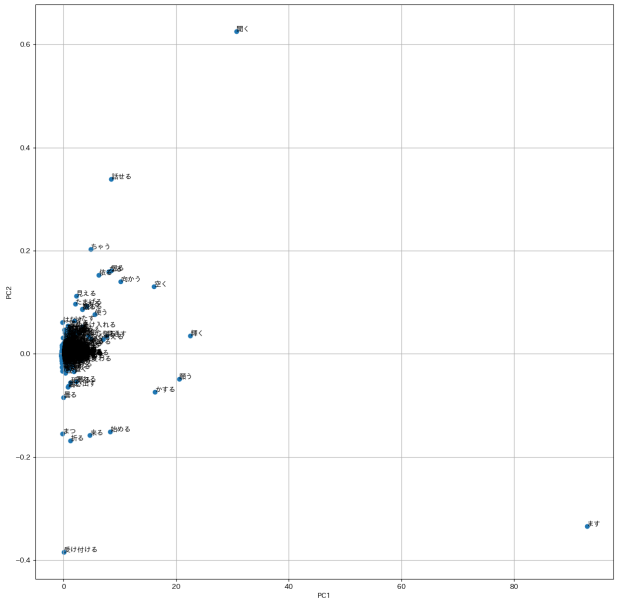

図 6.6. 主成分得点

日内で主成分分析したときと似たような位置にある単語が多い。

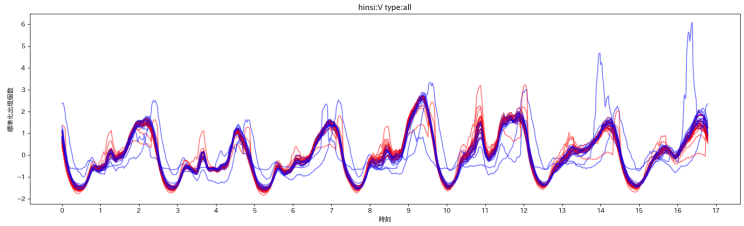

図 6.7. 主成分得点上位 10 単語の時系列変化

日内で夜遅くに多くなる変化（第 2 主成分得点が大きい）が、そのまま週内での第 2 主成分となっている。

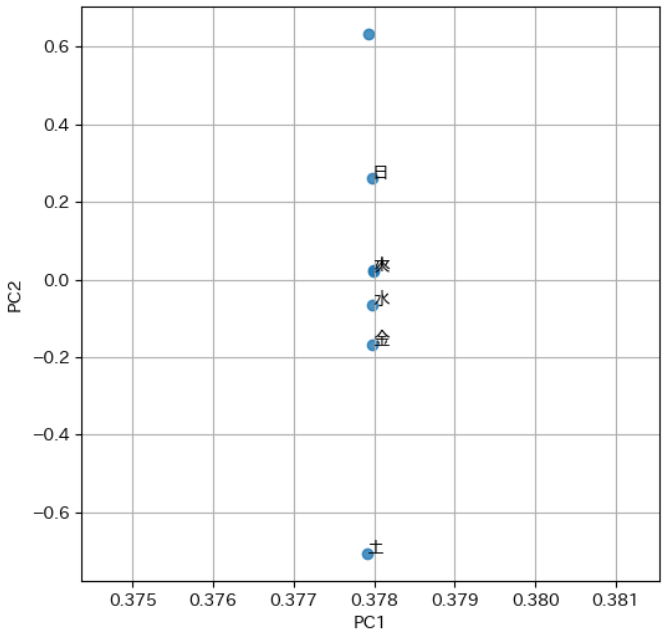

図 6.8. 主成分に対する観測変数の寄与度

PC1 は全体から満遍なく寄与を受ける。PC2 は月曜と土曜から強く寄与を受ける。

6.3.3　形容詞（日変動）

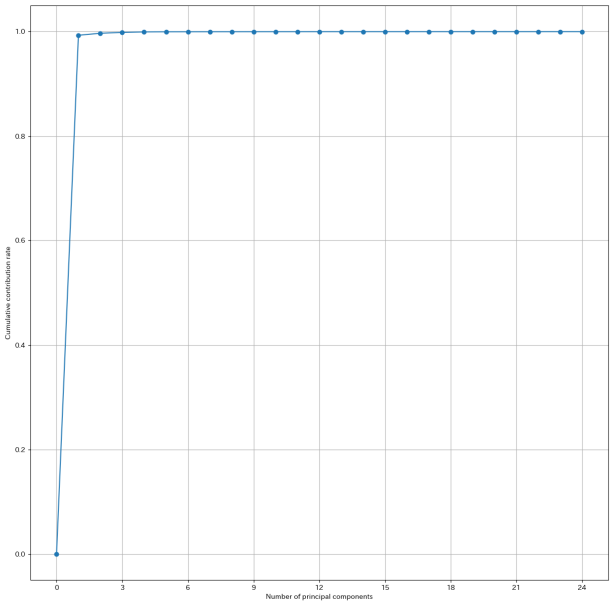

図 6.9. 累積寄与率

ほぼ第一主成分のみで説明できてしまっている。

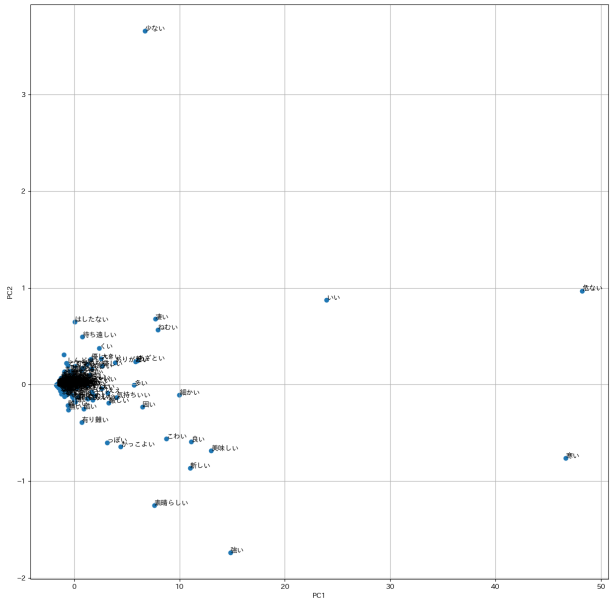

図 6.10. 主成分得点

解釈しづらい。

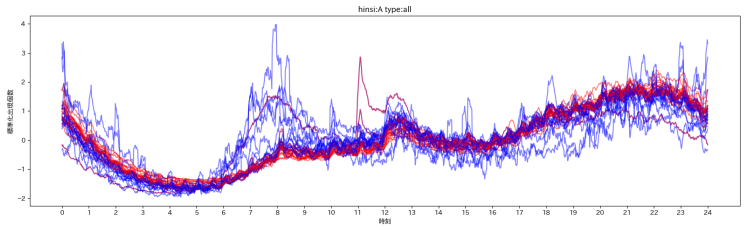

図 6.11. 主成分得点上位 10 単語の時系列変化

第 2 主成分得点が高い単語はは朝に多く、第 1 主成分得点が高い単語は夜に多い。

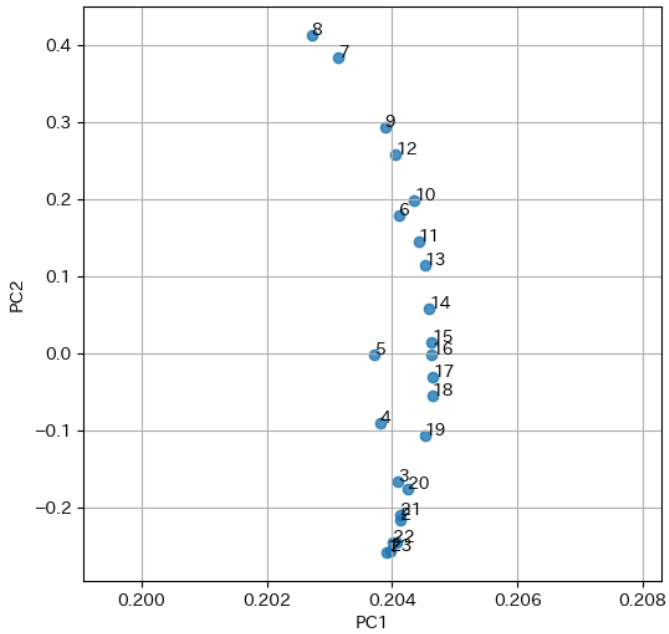

図 6.12. 主成分に対する観測変数の寄与度

PC1 には。昼ー夕方と朝あたりのデータが特に寄与している。PC2 は朝と夜に強く影響を受ける。

6.3.4 形容詞（週変動）

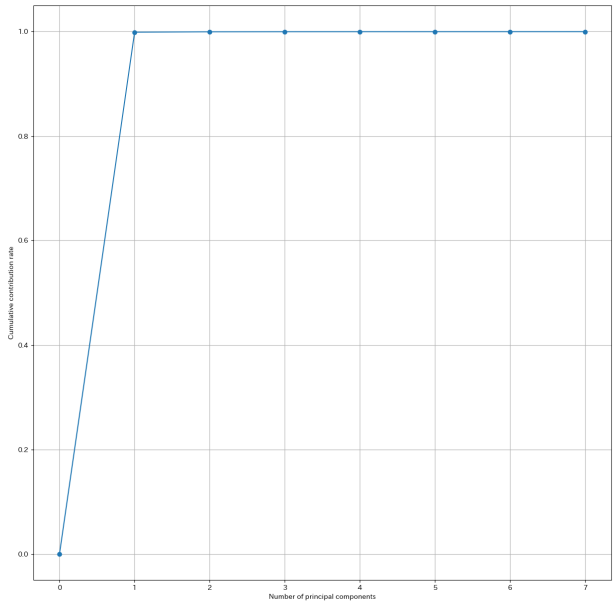

図 6.13. 累積寄与率

ほぼ第一主成分のみで説明できてしまっている。

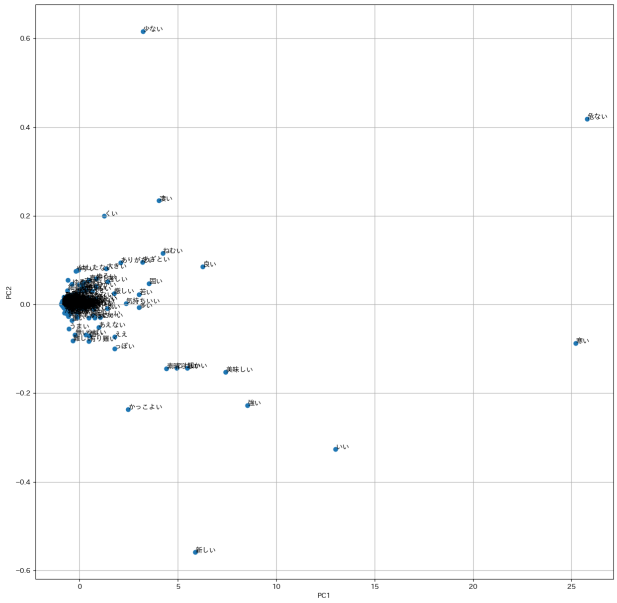

図 6.14. 主成分得点

日内変動のときと似たような配置になっている。

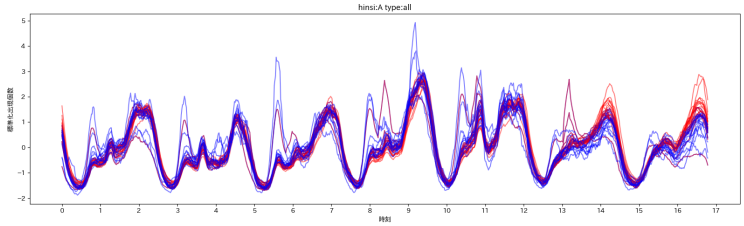

図 6.15. 主成分得点上位 10 単語の時系列変化

第 2 主成分得点が大きい単語は、第 1 主成分得点が大きい単語に比べて平日内での変化が激しい。

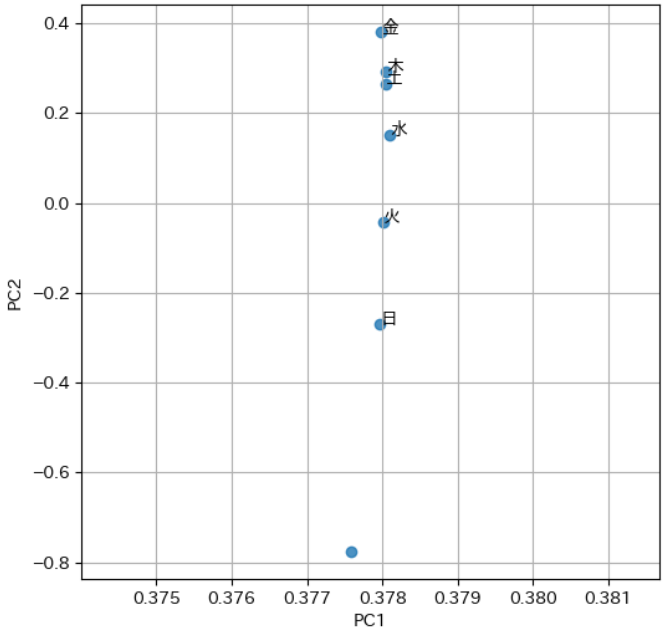

図 6.16. 主成分に対する観測変数の寄与度

PC1 は全体から満遍なく寄与を受ける。PC2 は月曜と金曜から強く寄与を受ける。

6.4 考察

日内変動と週内変動の主成分得点を散布図にすると、単語の配置が似ていることから、日内変動が週内変動に大きく寄与していると考えられる。これは先行研究と一致する。

第 7 章

まとめ

7.1　結論

クラスタリングにおいて、感情、行動表現の時系列変動に周期性が認められた。客観的な分析表現は比較的朝から昼に多く、主観的な感情表現は比較的夜に多いことがわかった。また、状態を表す動作表現は比較的朝から昼に多いく、瞬間的な変化を表す動作表現は比較的夜に多いことがわかった。しかし、反応などのパラメータとの関係や主成分の特徴を捉えることはできなかった。

7.2　課題

全体的なクラスタリングと主成分分析の精度を上げ、単語の種類数を増やし、特異な変化をする単語の特徴を調べるために、さらに長い期間の間、より多くのデータを収集する必要がある。

また、挨拶メッセージの除去など、前処理を工夫する必要がある。

第 8 章

参考文献

[1] Fabon Dzogang, Stafford Lightman, Nello Cristianini (2018) Diurnal variations of psychometric indicators in Twitter content, PLOS ONE(Open Access Journal)
[2] LIWC, http://www.liwc.net/
[3] 鳥海不二夫、榊剛史、吉田光男 (2020) ソーシャルメディアを用いた新型コロナ禍における感情変化の分析、人工知能学会論文誌
[4] サンプリングされたストリーム-TwitterAPI,https://developer.twitter.com/en/docs/twitter-api/tweets/sampled-stream/api-reference/get-tweets-sample-stream
[5] 工藤拓 (2013) MeCab, http://taku910.github.io/mecab/