特別研究報告題目

Word2Vecとクラスタ分析を利用した感情推定モデル

Emotion Estimation Model using Word2vec and Clustering Analysis

指導教員　主査 出口 利憲 教授
副査 田島 孝治 准教授

岐阜工業高等専門学校　専攻科　先端開発融合専攻

2023Y14 土田陸斗

令和7年 (2025年) 1月29日提出

Abstract

This study proposes a novel emotion estimation model based on Plutchik's eight basic emotions, utilizing Word2vec and cluster analysis. Unlike conventional methods using polarity dictionaries or machine learning, our approach requires no training data and enables multi-dimensional emotion representation.

The model employs Word2vec for word vectorization and a constrained K-means method for clustering. By fixing the centers of emotion clusters and updating only the centers of other clusters, we effectively classified words related and unrelated to emotions. The elbow method determined the optimal number of clusters: 8 for emotions and 20 for others.

We evaluated the model using song lyrics with pronounced emotions, plotting results on Plutchik's emotion wheel. Initial bias towards the "surprise"cluster was mitigated by adjusting cluster centers and introducing new non-emotion clusters.

Further improvements to the clustering algorithm included the application of K-means++ and Fuzzy C-means (FCM). K-means++ optimizes the initial centroid selection for non-emotion clusters, while FCM allows words to belong to multiple clusters, addressing multi-emotion words and improving the accuracy of analysis for texts with fewer words.

The model demonstrated the ability to capture complex emotional nuances in text, offering a flexible and multifaceted approach to emotion estimation. Future work will focus on processing negations and antonyms, considering context, and establishing quantitative evaluation methods.

This research contributes to the field of natural language processing by providing a new perspective on emotion analysis tasks, paving the way for more nuanced and comprehensive emotion estimation in various applications.

目 次

Abstract

第 1 章　序論 **1**

第 2 章　提案手法 **3**

2.1　概要 3
2.2　Word2Vec による単語のベクトル化 3
2.2.1　Word2Vec について 3
2.3　制約付き K-means 法によるクラスタリング 3
2.3.1　K-means 法について 3
2.3.2　感情クラスタと他クラスタの定義 4
2.3.3　エルボー法 4
2.3.4　シルエット係数 4
2.3.5　クラスタリングアルゴリズムの詳細 5
2.4　感情辞書の作成 6
2.4.1　辞書のデータ構造 6
2.5　感情分析の実行 6
2.5.1　入力テキストの前処理 6
2.5.2　感情値の計算 7

第 3 章　実験 **8**

3.1　実験設定 8
3.1.1　使用データセット 8
3.1.2　使用する Word2Vec モデル 8
3.1.3　評価指標 8
3.2　基本的な結果 9
3.2.1　クラスタ数の決定 9
3.2.2　初期結果 9

第 4 章　改善 **11**

4.1　モデルの改善 11
4.1.1　驚きクラスタの調整 11
4.1.2　助動詞処理の改善 11
4.1.3　改善後の出力 11
4.1.4　歌詞以外のテキストの入力 12

4.2 クラスタリングアルゴリズムの改善 12
4.2.1 K-means++の使用 12
4.2.2 Fuzzy C-means の使用 13

第 5 章 考察 **15**
5.1 基本モデルの性能評価 15
5.1.1 全体の評価 15
5.1.2 モデルの問題点 15
5.2 K-means++との比較分析 15
5.3 FCM との比較分析 16

第 6 章 結論と今後の課題 **17**
6.1 研究成果のまとめ 17
6.2 今後の研究課題 17

第1章　序論

自然言語処理技術は、ChatGPT を筆頭に急速な発展を遂げており、音声対話システムや検索エンジン等、私たちの生活に欠かせない幅広い分野で応用されている。この技術の進歩に伴い、テキストから感情を推定する感情分析タスクの重要性も増している。感情分析は、SNS の分析、顧客レビューの評価、マーケティング戦略の立案など、多岐にわたる領域で活用されている。

従来の感情分析手法として、単語に –1 から 1 の極性を割り振った辞書を用いて文章のポジティブおよびネガティブ度合を数値化する方法が一般的であった [1]。しかし、人間の感情は極めて複雑であり、一次元の数値では厳密な感情を完全に捉えることは困難である。また、機械学習を用いた手法も提案されているが、大量の教師データが必要であり、モデルの柔軟な変更が困難であるなどの問題がある [2]。

水間ら [3] は、Word2Vec を用いて文章をベクトル化し、8 つの基本感情ベクトルとの各コサイン距離をパラメータとする手法を提案した。しかし、この手法では感情に関係のない単語の寄与が大きく、人間が感じた感情と異なる結果が得られることがあった。安達ら [4] は 10 種類の感情についての感情語辞書を作成し、1 秒以内に 10,000 文といった高速な分析を可能にしたが、モデルの柔軟性に欠け、感情カテゴリの追加、削除、修正が大きな課題になっている。

本研究室における先行研究 [5] により、Word2Vec によりベクトル化した単語に対し、K-means 法を適用することで、単語の意味を考慮したクラスタリングを実現できることが分かっている。また、奥出ら [6] は、学習した Word2Vec モデルから顔文字のベクトルを算出し、K-means 法によるクラスタリングを行った。これにより、顔文字間の類似度に基づいて、顔文字の感情を分析する手法が有効である可能性が示されている。これらの結果から、テキストベースの感情分析において、Word2Vec モデルに K-means 法を適用することで、各単語の感情に基づいたクラスタリングが実現できるという仮説を立てた。

本研究では、Word2Vec とクラスタ分析を用いて、プルチックが提唱した 8 つの基本感情 [7] に基づいた新しい感情推定モデルの構築を目指す。具体的には、以下の目的を設定する。

1. プルチックの 8 つの基本感情に基づいた多次元的な感情推定モデルの開発
2. 教師データを必要としない柔軟性の高い感情分析手法の確立
3. 感情に関連する単語と関連しない単語を効果的に分類する新たなクラスタリング手法の開発

本研究により、より柔軟で多面的な感情推定が可能となり、自然言語処理における感情分析タスクの精度向上に貢献することが期待される。

第2章　提案手法

2.1　概要

本研究では、Word2Vec とクラスタ分析を組み合わせた感情推定モデルを提案する。この手法は教師データを必要とせず、複数の感情を同時に考慮できる点で既存手法と異なる。提案手法は以下の 4 つの主要ステップから成る：

1. Word2Vec による単語のベクトル化
2. 制約付き K-means 法によるクラスタリング
3. 感情辞書の作成
4. 感情分析の実行

2.2　Word2Vec による単語のベクトル化

2.2.1　Word2Vec について

Word2Vec は、自然言語処理における単語の分散表現手法として広く認知されている。この手法は、大規模なテキストコーパスから単語の意味的関係性を学習し、それを低次元の密ベクトルとして表現する。Word2Vec の主要な特徴は以下の通りである。

- **意味的類似性の捕捉**: 意味的に類似した単語は、ベクトル空間上で近接して配置される。
- **類推タスクの実現**: 単語間の関係性を、ベクトルの加減算で表現できる（例：「王」−「男」＋「女」≈「女王」）。
- **次元削減**: 高次元の単語表現を低次元のベクトルに圧縮し、計算効率を向上させる。

2.3　制約付き K-means 法によるクラスタリング

2.3.1　K-means 法について

K-means 法は、データを k 個のクラスタに分類するクラスタリング手法の一つである。データの重心を求めることで分類を行い、比較的大規模なデータセットにも適用可能で幅広い分野で利用されている。次に、K-means 法の手順を示す。

1. クラスタ数 k を決定する

2. k 個の初期重心をランダムに設定する

3. 各データポイントを最も近い重心のクラスタに割り当てる

4. 各クラスタの新しい重心を計算する

5. 収束するまで手順 3 と手順 4 を繰り返す

2.3.2 感情クラスタと他クラスタの定義

本研究では、感情語のクラスタを感情クラスタ、その他の感情に関係のない単語のクラスタを他クラスタと呼ぶ。また、それぞれのクラスタ数については後述する、エルボー法とシルエット係数を用いて決定する。

2.3.3 エルボー法

エルボー法 [8] は、クラスタリング分析において適切なクラスタ数を決定するための一般的な手法である。この方法は、クラスタ内平方和（WCSS: Within-Cluster Sum of Squares）を用いてクラスタの品質を評価する。WCSS は以下の式で定義される。

$$WCSS = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{2.1}$$

ここで、k はクラスタ数、C_i は i 番目のクラスタ、x はデータポイント、μ_i は i 番目クラスタの重心である。エルボー法の手順は以下の通りである。

1. 異なるクラスタ数 k に対してクラスタリングを実行する（例：$k = 1$ から $k = 10$ まで）。

2. 各 k に対して WCSS を計算する。

3. クラスタ数 k に対する WCSS のグラフを作成する。

4. グラフ上で「エルボー」（急激な曲がり）が見られる点を特定する。

この「エルボー」の位置が最適なクラスタ数を示唆する。これは、この点以降にクラスタ数を増やしても、WCSS の減少が緩やかになることを意味する。

2.3.4 シルエット係数

シルエット係数 [9] は、クラスタリングの品質を評価するための指標の一つである。この係数は、各データポイントがどの程度適切にクラスタリングされているかを数値化し、クラスタの分離度と凝集度を同時に評価する。シルエット係数は以下の式で定義される。

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2.2}$$

ここで、

- $a(i)$: データポイント i と同じクラスタ内の他のすべてのデータポイントとの平均距離（クラスタ内凝集度）
- $b(i)$: データポイント i と最も近い他のクラスタのすべてのデータポイントとの平均距離（クラスタ間分離度）

であり、特徴として次のようなものが挙げられる。

- 値の範囲は -1 から 1 までである。
- 高い値（1 に近い）は、データポイントが自身のクラスタによく適合し、他のクラスタとは明確に分離されていることを示す。
- 低い値（-1 に近い）は、データポイントが誤ったクラスタに割り当てられている可能性を示唆する。
- 0 に近い値は、データポイントがクラスタの境界付近に位置していることを意味する。

2.3.5 クラスタリングアルゴリズムの詳細

本研究で使用するクラスタリングアルゴリズムは、K-means クラスタリングをベースとしたものである。アルゴリズムの全体的な流れは Figure 2.1に示されている。以下、アルゴリズムの各ステップについて詳述する。

1. **初期重心の割当**
    - 感情クラスタの初期重心: 各感情語を設定
    - 他クラスタの初期重心: ランダムに選ばれた単語を設定
2. **クラスタリングの実行**
    - 通常の K-means 法と同様に、各単語を最も近いクラスタ重心のクラスタに分類
3. **重心の再計算**
    - 他クラスタの重心のみを更新
    - 感情クラスタの重心は固定したまま維持

このプロセスを収束するまで繰り返す。

このアルゴリズムの特徴は、感情クラスタの重心を固定することで、感情語を中心としたクラスタ形成を可能にしている点である。これにより、各感情に関連する単語群を効果的に抽出することができる。

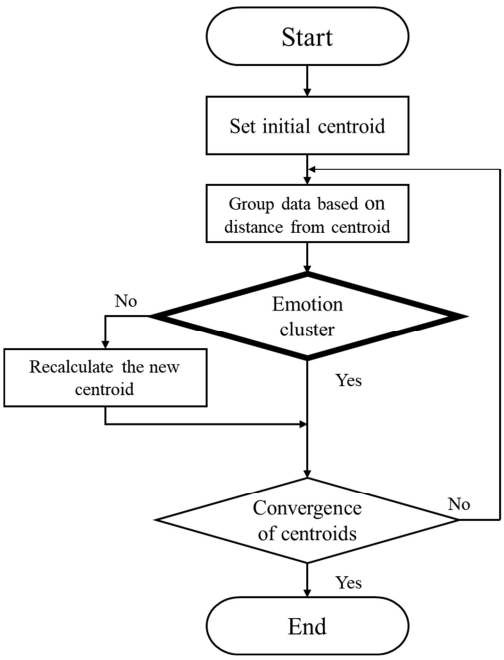

Figure 2.1: Flow chart of clustering.

Table 2.1: Structure of dictionary.

Word	**Emotion Label**	**Weight**
Each word	Cluster Type	COS Similarity to the centroid

2.4 感情辞書の作成

2.4.1 辞書のデータ構造

本研究では、クラスタリング結果に基づいて感情分析用の辞書を構築した。この辞書は、単語、感情ラベル、重みの 3 要素から構成される。Table 2.1に辞書のデータ構造を示す。感情ラベルは単語が所属するクラスタの種類を表し、重みは単語と所属クラスタの重心とのコサイン類似度により算出される。

2.5 感情分析の実行

2.5.1 入力テキストの前処理

感情の分析を実行するにあたり、MeCab[10] を使用して入力テキストの前処理を行う。MeCab とは日本語に対応したオープンソースの形態素解析エンジンであり、文書から品詞、単語などの情報をもとに、言語の最小単位を指す形態素に分割する。

2.5.2　感情値の計算

分かち書きされた文章を参照し、各形態素について、作成した辞書と比較する。辞書に含まれる単語とラベルから、各感情の重みを加算し、その合計をプロットする。

第3章　実験

3.1　実験設定

3.1.1　使用データセット

実験に使用するデータセットとして、感情が顕著に表れている曲を採用した。ここでは、特に楽曲としての表現が明瞭で、かつ客観的な評価がしやすい5つの感情に限定した。「喜び」「悲しみ」「怒り」「嫌悪」「恐怖」といった感情は、歌詞やメロディにおいて極端な表現がされやすく、解釈のばらつきが少ないため、解析結果の信頼性を担保しやすいと考えた。Table 3.1は実験に使用した5つの曲を表にまとめたものであり、主な感情、曲名、アーティスト名、歌詞の形態素数が含まれている。

3.1.2　使用する Word2Vec モデル

本研究では、使用する Word2Vec モデルに学習済み日本語 Wikipedia エンティティベクトル [11] の 100 次元のものを採用した。これは東北大学の乾研究室で作成され、幅広い分野の単語を含む大規模なコーパスに基づいている。

3.1.3　評価指標

複数の感情を分析するモデルについて、定量的評価を行うのは困難である。したがって、本研究では入力した曲の感情と、出力結果を人間の感覚により定性的に評価した。

Table 3.1: List of Songs for Analysis.

Type of Song	**Song Title**	**Artist**	**Number of Morphemes**
Happy Song	ハピネス	AI	396
Sad Song	ドライフラワー	優里	299
Angry Song	うっせえわ	Ado	323
Dislike Song	五月の蠅	RADWIMPS	504
Scary Song	呪い	山崎ハコ	184

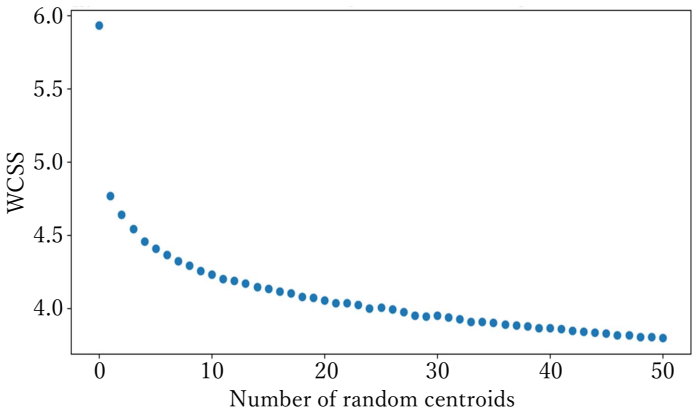

Figure 3.1: WCSS in clusters.

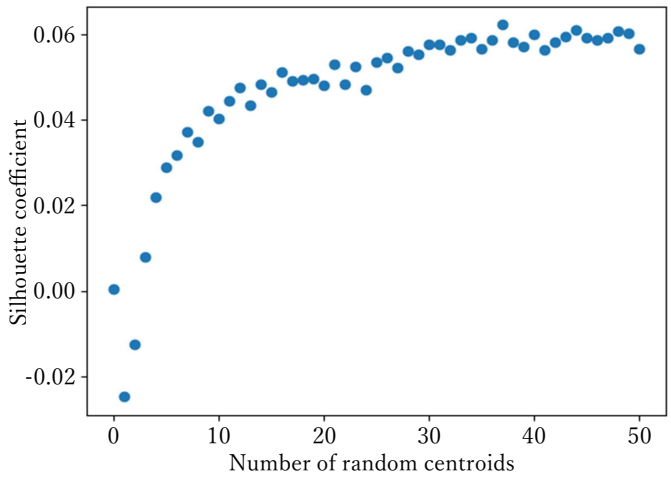

Figure 3.2: Silhouette coefficients.

3.2 基本的な結果

3.2.1 クラスタ数の決定

Figure 3.1はエルボー法によるクラスタ数決定のために WCSS をプロットしたものである。横軸がクラスタ数、縦軸が WCSS を表す。また、Figure 3.2は同様にシルエット係数をプロットしたものであり、横軸がクラスタ数、縦軸がシルエット係数になっている。この分析により WCSS が 4.1 以下、シルエット係数が 0.045 以上に収束する値として、他クラスタの数は 20 に設定した

3.2.2 初期結果

Figure 3.3に 5 つの曲を入力した時の出力を示す。これはプルチックの 8 つの基本感情の輪に基づき、鶏頭図としてプロットしたものである。全体的な傾向として「驚き」クラスタへの偏りが顕著であり、ほとんどの楽曲で「驚き」の値が高くなっている。

「驚き」のクラスタに分類された単語をのぞいて考えると、Happy song, Angry song は分

Figure 3.3: First result.

類数が最多のものと曲の感情が一致しているが、他の３つは適切に分類できていなかった。

第4章　改善

4.1　モデルの改善

4.1.1　驚きクラスタの調整

驚きクラスタへの誤った分類を踏まえ、初期重心に採用する単語の再検討を行った。実験を通じて、驚きクラスタの初期重心を「驚き」から「驚愕」へと変更し、より適切な分類が行われるよう調整した。

4.1.2　助動詞処理の改善

入力されたテキストの各形態素の分類先を見たところ、文章によく出現する、「です」「ます」のような助動詞が誤って驚きクラスタにクラスタリングされていることが分かった。そこで、これらの助動詞の誤分類への対策として、重心を更新しない新たなクラスタを作成し、その重心を「です」に設定した。この措置により、助動詞「です」「ます」などが誤って驚きクラスタに分類されることを防ぐことが可能となった。

4.1.3　改善後の出力

Figure 4.1は、上記改善後の出力を示している。驚きクラスタへの過度な偏りが大幅に減少し、各曲の感情により適合した分析結果が得られている。この結果から、改善策が有効であり、モデルの性能が向上したことが確認できる。また、形態素の数が多い Dislike song では複数の感情を読み取っているのに対し、形態素の数が少ない Scary song では「喜び」と「悲しみ」以外の感情は読み取られていない。

Figure 4.1: Improved result.

Table 4.1: Key Scenes in the Story of Melos.

Parts	Scene and Emotion
1	Melos, enraged by the king's cruelty, decides to confront him.
2	Melos, filled with hope, prepares for his sister's wedding.
3	Exhausted, Melos falls asleep after the wedding.
4	In a dream, Melos feels anxiety and doubt about his fate.
5	Melos runs desperately to save his friend, trusting he'll make it.
6	Melos saves his friend, feeling the strength of trust and friendship.

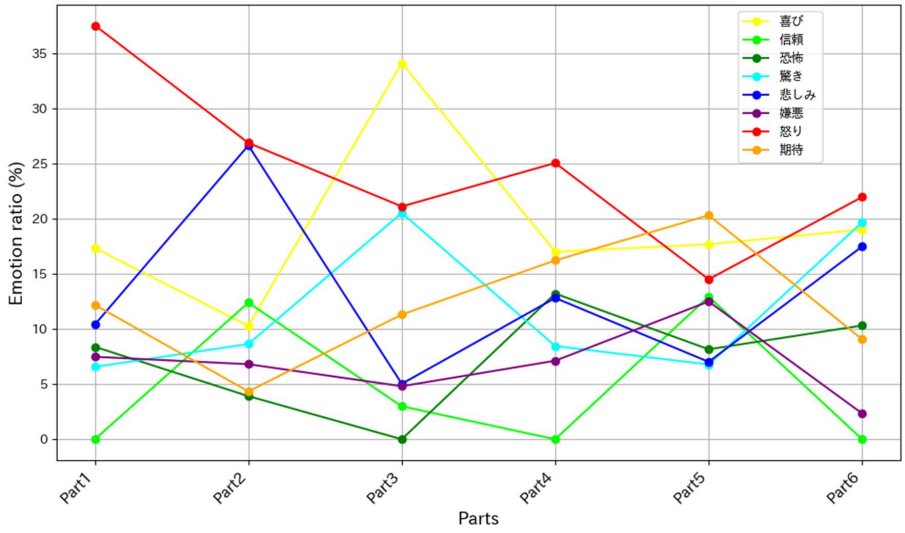

Figure 4.2: Results with "走れメロス".

4.1.4 歌詞以外のテキストの入力

改善後のモデルについて、歌詞以外のテキストへの対応を見るべく、物語のテキストを入力した。Figure 4.2は太宰治の走れメロスを、物語の構成で 6 つに分割し、それぞれのセクションについて分析して、感情の推移をプロットしたものである。ここで、Table 4.1は 6 つに分割した各セクションの主な特徴である。

4.2 クラスタリングアルゴリズムの改善

4.2.1 K-means++の使用

K-means++[12] は、従来の K-means アルゴリズムの初期値選択問題を解決するために提案された改良手法である。本手法は、初期クラスタ中心をデータ点間の距離に基づいて確率的に決定することで、より効果的なクラスタリングを実現する。K-means++のアルゴリズムは以下の手順で実行される。

1. データセットからランダムに 1 点を選択し、最初のクラスタ中心とする。

2. 各データ点に対し、既存のクラスタ中心との最小距離 $D(x)$ を計算する。

Figure 4.3: Results with K-means++.

3. 重み付き確率分布 $\frac{D(x)^2}{\sum D(x)^2}$ に基づいて、新たなクラスタ中心を選択する。

4. 目的のクラスタ数 k に達するまで、手順 2 と 3 を繰り返す。

5. 得られた初期クラスタ中心を用いて、通常の K-means アルゴリズムを実行する。

この手法により、初期クラスタ中心が互いに離れて配置される確率が高くなり、局所解に陥るリスクが軽減される。結果として、K-means++は K-means と比較して、より安定した収束性能と高品質なクラスタリング結果を提供する。

また、実験でこれを実装する際、感情クラスタの初期重心は考慮せず、K-means++で決定した初期重心に後から結合する形を採用した。

Figure 4.3は K-means++を使用して他クラスタの初期重心を決定した時の出力である。入力テキストは先述した実験と同じものを使用している。図からわかるように、今回使用した歌詞の分析結果では、改善前の K-means の結果と大きな違いは見られなかった。

4.2.2 Fuzzy C-means の使用

Fuzzy C-means (FCM)[13] は、データ点が複数のクラスタに部分的に所属することを許容する柔軟なクラスタリングアルゴリズムである。従来のハードクラスタリング手法とは異なり、FCM では各データ点がすべてのクラスタに対して 0 から 1 の範囲の所属度を持つ。FCM アルゴリズムは以下の目的関数を最小化することで最適なクラスタリングを得る:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \left| \mathbf{x}_i - \mathbf{c}_j \right|^2 \tag{4.1}$$

ここで、N はデータ点の数、C はクラスタ数、m はファジネスパラメータ（通常 $m > 1$）、u_{ij} はデータ点 i のクラスタ j への所属度、$\mathbf{x}_i$ は i 番目のデータ点、$\mathbf{c}_j$ は j 番目のクラスタ中心である。FCM アルゴリズムは以下の手順で実行される:

1. クラスタ中心 $\mathbf{c}_j$ を初期化する。

2. 各データ点 i の各クラスタ j への所属度 u_{ij} を更新する:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{|\mathbf{x}_i - \mathbf{c}_j|}{|\mathbf{x}_i - \mathbf{c}_k|} \right)^{\frac{2}{m-1}}} \tag{4.2}$$

Figure 4.4: Results with Fuzzy C-means.

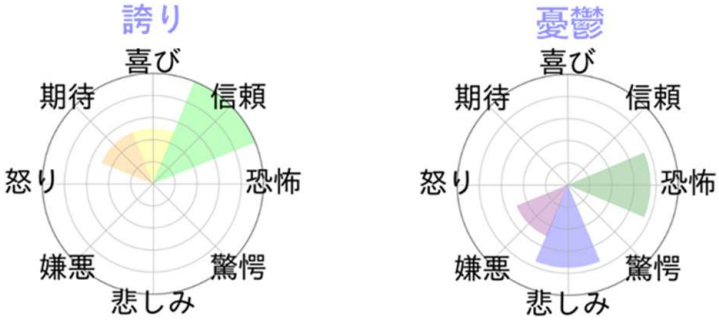

Figure 4.5: Result for a single word using FCM.

3. クラスタ中心 $\mathbf{c}_j$ を更新する:

$$\mathbf{c}_j = \frac{\sum_{i=1}^{N} u_{ij}^m \mathbf{x}_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{4.3}$$

4. 収束条件を満たすまで手順 2 と 3 を繰り返す。

FCM は従来の K-means と比較して、データの曖昧性や不確実性をより適切に扱うことができる。

Figure 4.4に、FCM による結果を示す。これについても、K-means++の出力結果と同様、改善前のものとの間に大きな差異は見られなかった。ただし、FCM により作成した辞書を用いて「誇り」「憂鬱」という 1 つの単語に対しても分析を行った。Figure 4.5に示すように、この単語は複数の感情が混在していることがわかる。つまり、単語レベルの感情の複雑性を捉えることが可能になった。

第5章　考察

5.1　基本モデルの性能評価

5.1.1　全体の評価

Figure 4.1を見てわかるように、各曲の主要な感情を読み取れており、人間の感覚と概ね一致する精度での分析ができた。

Figure 4.2では、物語の進行に伴い「怒り」の割合が減少し、「期待」の割合が増加しており、物語の大筋を適切に捉えていることが分かる。特に、王の暴政にメロスが激怒するPart1では「怒り」の割合が、メロスの妹が結婚式を挙げるPart3では「喜び」の割合が高くなっており、各場面の感情を正確に読み取れていることが示されている。最後の場面で怒りの割合が突出しているのは、文中に「殴る」という単語が頻出したためと考えられる。

5.1.2　モデルの問題点

Sad songで「喜び」の感情値が高くなっていたり、Scary songで「恐れ」の感情値が0になったりしており、一部人間の感覚と乖離した部分がある。これらの原因として、主に次の3つが考えられる。

一つ目は辞書を使い、単語ごとの感情値を計算する手法をとっているため、文脈に依存した感情を読み取れない点である。例えば、「私は幸せじゃない」という文では、「幸せ」という単語自体は肯定的な感情を持つが、文全体としては否定的な意味になる。このような文脈依存の感情を現在の手法では正確に捉えることができない。

二つ目の原因として考えられるのは、歌詞の中に含まれる比喩表現や隠喩、皮肉などの修辞的表現を適切に解釈できない点である。例えば、「心が冷めた」という表現は、文字通りの意味ではなく、感情の冷たさや無関心を表現しているが、現在の手法ではこのような比喩的な表現の真の感情を理解することが難しい。

三つ目の原因として、辞書作成、クラスタリングの過程における分類の不適切性が挙げられる。本手法の利点は人的介入を最小化できる点にあるが、これ自体が、一部の語彙を人間の感覚から逸脱した形で分類してしまう可能性を内包している。

5.2　K-means++との比較分析

先述の通り、K-means++を適用した場合でも、結果には大きな差異は見られなかった。Figure 5.1は、クラスタ重心の遷移を示したものであり、100次元のベクトルデータを主成分分析により2次元へ次元削減した後にプロットしたものである。なお、この図の作成に

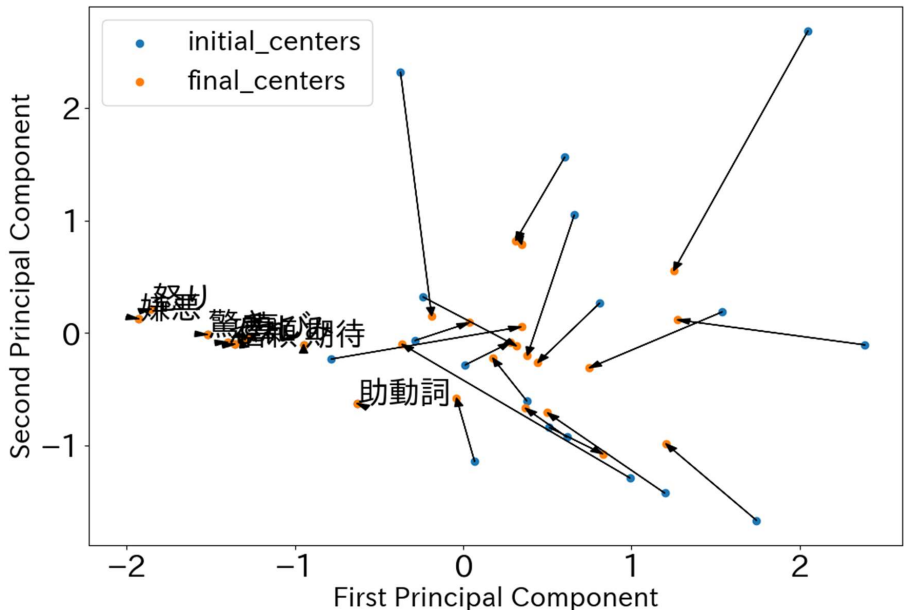

Figure 5.1: Movement of centroids.

際して、他クラスタの初期重心は K-means++ではなくランダムに設定している。

図から示されるように、感情クラスタは他クラスタと比較して密集している。このことから、初期クラスタ重心が互いに離れる傾向にある K-means++アルゴリズムを使用した場合でも、改善されるのは感情クラスタから離れた他クラスタであり、感情クラスタに対する影響は限定的であることが示唆される。

また、「です」「ます」などの助動詞が感情語群に近い位置に分布しており、これを調整することでモデルの精度が向上した。このことから、モデルの精度向上には、感情クラスタ近傍の他クラスタに対する適切な調整が重要であると考えられる。

5.3 FCM との比較分析

先述の通り、Fuzzy C-means 法（FCM）を適用した結果も、従来の手法と比較して顕著な差異は観察されなかった。加えて、Figure 4.1と Figure 4.4の比較からも明らかなように、FCM を用いた場合、各曲における主要な感情スコアが全体的に低くなっていることが確認された。これは、FCM がソフトクラスタリングを採用しているため、各単語が複数の感情クラスに属することにより、主要な感情以外のクラスに属する度合いが相対的に増加した結果であると考えられる。この特性により、FCM を用いた感情分析モデルは、長文の分析には適していない可能性が示唆される。

今後の研究では、FCM の適用範囲や効果をより詳細に検討し、異なる文章構造に対する適用性を明らかにする必要があると考えられる。

第6章　結論と今後の課題

6.1　研究成果のまとめ

本研究では、プルチックの8つの基本感情に基づく新しい感情推定モデルを提案した。本モデルは、従来の辞書ベースまたは機械学習の手法と異なり、Word2Vec による単語ベクトル化とクラスタリングアルゴリズムを組み合わせ、教師データを必要としない柔軟な感情分析を可能にした。特に、感情クラスタの初期重心を固定することで、感情に関連する単語の分類精度を向上させた。実験結果では、複数の楽曲を用いた感情分析において、プルチックの感情輪に基づく感情の分布を視覚的に確認した。また、従来の K-means 法に加え、K-means++および Fuzzy C-means (FCM) を用いたクラスタリング手法の改善を試みた。特に、K-means++では初期クラスタ中心の選定を最適化し、FCM では単語が複数の感情クラスタに所属する可能性を考慮した。これらの成果から、本研究は自然言語処理分野における感情分析タスクに新たな視点を提供し、感情の多次元的な表現および分析精度の向上に寄与するものである。

6.2　今後の研究課題

本研究により、感情分析における多次元的な推定モデルの有効性を示すことができたが、いくつかの課題が残されている。これらの課題に対処することで、モデルのさらなる精度向上が期待される。

まず、否定語や対義語、文脈依存の感情処理に対する改善が必要である。現行のモデルでは、単語単位での感情推定が主であり、文全体の文脈や否定表現が十分に考慮されていない。例えば、「幸せではない」や「不満足」という表現において、否定語によって本来の感情が反転するケースがあるが、これらは現在の手法では正確に処理できない。

次に、定量的な評価基準の確立が今後の重要な課題となる。本研究では、主に定性的な評価を行ったが、モデルの客観的な性能を評価するためには、より具体的で信頼性のある定量的評価指標が必要である。特に、感情分類の精度や再現率、F 値などの従来の評価指標に加え、感情の多次元性や複雑性を反映する新たなメトリクスの開発が求められる。また、他の感情分析モデルとの比較を行うための標準データセットの整備や、長文テキストや比喩表現を含むデータに対する評価手法の確立も重要である。

謝辞

本研究を進めるにあたり、多大なるご指導とご助言を賜りました出口利憲先生に、心より感謝申し上げます。

参考文献

[1] 東北大学 乾・鈴木研究室：日本語評価極性辞書,
https://www.cl.ecei.tohoku.ac.jp/Open_Resources-Japanese_Sentiment_Polarity_Dictionary.html
（2024 年 9 月 10 日アクセス）

[2] Boiy, E., and Moens, M. (2009). A machine learning approach to sentiment analysis in multilingual Web texts. Information Retrieval, 12, 526–558.
https://doi.org/10.1007/s10791-008-9070-z

[3] 水間大資 他 3 名：Word2Vec を用いた文章の感情取得の研究，平成 29 年度電気・情報関係学会九州支部連合大会，2017.
https://www.jstage.jst.go.jp/article/jceeek/2017/0/2017_361/_pdf/-char/ja

[4] 安達由洋 他 4 名：感情語辞書を用いた日本語分の感情分析，可視化情報学会誌，2021.
https://www.jstage.jst.go.jp/article/jvs/41/161/41_21/_pdf/-char/ja

[5] 瀬尾慎介：K-means 法による次元削減を用いた文書分類，岐阜工業高等専門学校電気情報工学科卒業研究報告，2022.

[6] 奥出悠騎 他 3 名：Word2Vec を用いた顔文字感情分析手法におけるクラスタリングの有効性，情報処理学会代第 81 回全国大会，pp. 123-124，2019.
https://ipsj.ixsq.nii.ac.jp/ej/?action=repository_uri&item_id=196785&file_id=1&file_n

[7] Plutchik, R., and Kellerman, H. (1980). Theories of Emotion. Academic Press.

[8] Kodinariya, T. M., and Makwana, P. R. (2013). Review on Determining of Cluster in K-means Clustering. International Journal of Advance Research in Computer Science and Management Studies, 1(6), 90-95.
https://www.researchgate.net/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering
(アクセス日: 2023 年 10 月 26 日)

[9] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

[10] Kudo, T., and Matsumoto, Y. (2003). Fast methods for kernel-based text analysis. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1 (pp. 24-31).

[11] 鈴木 正敏. 日本語 Wikipedia エンティティベクトル.
http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/
（2023 年 10 月 7 日アクセス）

[12] Arthur, D., and Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms (pp. 1027-1035).

[13] Bezdek, J. C., Ehrlich, R., and Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 10(2-3), 191-203.