BERT に基づいた Russell 円環モデルの感情分析

古 泳欣　小林 一郎
お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻
{gu.yongxin,koba}@is.ocha.ac.jp

概要

Russell 円環モデルは，感情を「感情価（Valence）」と「覚醒度（Arousal）」の二軸で表現し，感情分布が円環状に配置されるを視覚化し，感情の認知構造を直感的に説明する枠組みとして広く利用される．本研究では，BERT モデルを用い，Russell 円環モデルに基づく感情分布を検証するために，次の 2 つの手法を提案する．一つ目は BERT の CLS トークンを特徴量として用いる方法（方法 1），二つ目は BERT の Attention ウェイトを利用して文全体の感情値を算出する方法（方法 2）である．これらの手法を用いて複数の実感情データセットを用いて実際の感情分布を分析し，この理論モデルの妥当性を検証するとともに，感情分布の実態に即した新たな知見を提供することを目的とする．

1　はじめに

近年，ソーシャルメディアやオンラインコンテンツの増加に伴い，テキストデータから感情を分析する重要性が高まっている．感情分析は，人間 – コンピュータインタラクションやロボット技術において自然で効果的な対話を実現する上で不可欠であり，感情を正確に認識することはより良い社会的サービスの提供に繋がる．

Russell の円環モデル [1] は, 感情を感情価（Valence）と覚醒度（Arousal）の 2 軸で表現し，感情分布を視覚化する枠組みとして広く使用されている．しかし，このモデルに基づく感情分布は主に経験的な主観に基づくものであり，実際の感情分布との整合性を検討する余地が存在する．Russell の円環モデル [1] は，感情を感情価（Valence）と覚醒度（Arousal）の 2 軸で表現し，感情が円環状に分布するという仮説を基に構築されたモデルである．このモデルは感情の認知構造を直感的に視覚化する手段として広く利用されているが，その分布が実際の感情データと一致するかどうかについては十分に検討されていない．本研究では，複数の大規模感情データセットを用いて感情分布を可視化し，理論モデルとの整合性を検証することで，感情モデルの改良や再構築の可能性を示唆する．

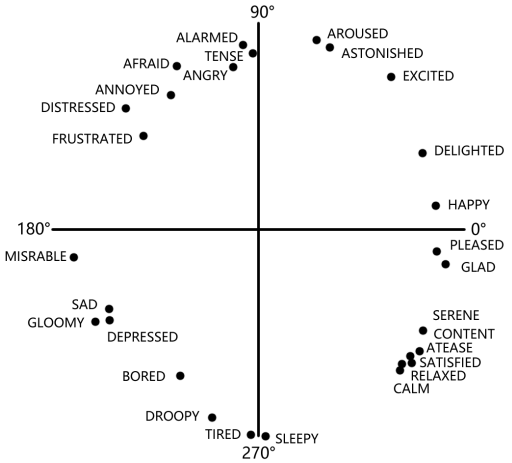

図 1　Russell 円環モデル：28 の感情語の円形座標

本研究では，BERT モデル [2] を用いて自然言語文章を対象にその中に表現される感情分布を検証し，Russell モデルに基づく分布との比較を通じて，最も合理的な感情分布を明らかにすることを目的としている．この検証により，感情分布における新たな視点を提供し，より現実的な感情理解を可能にすることを目指す．

2　関連研究

感情分析における感情価（Valence）および覚醒度（Arousal）（以下，VA 値）に関連する代表的なデータセットとして，XANEW[3] と EmoBank[4] が挙げられる．XANEW は 13,915 の英単語に対する VA 値を提供し，EmoBank は文単位の VA 値を含むデータセットであり，いずれも感情分析の評価において重要なリソースである．

VA 値の算出方法として，Horvat ら [5] は XANEW 辞書を用いて単語の感情スコアを集計しテキスト全体の VA 値を推定する手法を提案した．一方，近年

では BERT などの事前学習済み言語モデルを活用した深層学習アプローチが注目されている．Ito ら [6] は，BERT を用いた特徴抽出と SVR や DNN による回帰モデルを組み合わせ，少量のデータでも高精度な VA 値予測を実現した．

本研究では，BERT を用いた VA 値予測手法を採用し，Russell 円環モデルに基づく最適な感情分布の検証を目指す．

3 手法

本研究では，BERT モデルを基に，XANEW データセットを用い単語レベルの VA 値を回帰予測するモデルを構築した．さらに，このモデルを活用して，文レベルでの VA 値を算出するための二つの手法を提案する．本研究で提案する手法は，単語レベルの VA 値予測モデルを基盤に，それを用いて文レベルの VA 値を予測するものである．

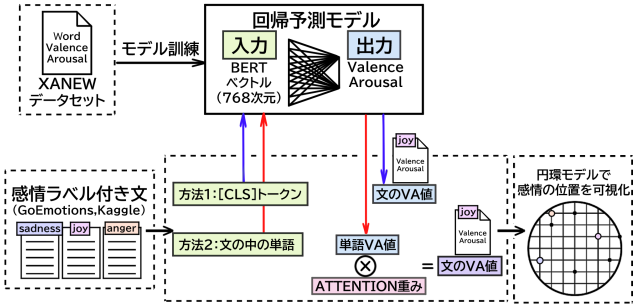

図 2 研究手法概要図

3.1 単語レベルの VA 値回帰モデルの構築

XANEW[3] は，ANEW[7] を拡張したデータセットで，感情価，覚醒度，支配度を 9 段階で評価した 1,034 語に加え，クラウドソーシングによって 12,000 語以上が追加されたリソースである．

本研究では，XANEW を基に BERT を用いて単語の特徴ベクトルから感情価と覚醒度を回帰予測するモデルを構築した．BERT モデルを使用して 768 次元の特徴ベクトルを抽出し，これを訓練データとして対応する VA 値を予測するニューラルネットワークを訓練した．このモデルにより，単語の特徴ベクトルから VA 値を精度高く推定することが可能となる．

3.2 入力文の準備と VA 値算出

使用する入力文は，GoEmotions[8] および Kaggle データセット [9] から取得した．

GoEmotions： Reddit から収集された 58,009 件のコメントで構成され，27 の感情カテゴリと neutral ラベルが注釈付けされたマルチラベル分類向けのコーパス．

Kaggle データセット： 20,000 件の Twitter メッセージを基にしたコーパスで，6 種類の基本感情（怒り，恐怖，喜び，愛，悲しみ，驚き）のラベルが付与されている．

これらの文に対し，以下の二つの方法で VA 値を算出した．算出結果は Russell 円環モデル上で可視化され，各感情の分布を分析することで，感情分布の特徴を明らかにすることを目的とする．

3.3 方法 1: [CLS] トークンを用いた予測

方法 1 では，文を BERT モデルに入力し，得られる [CLS] トークンの特徴ベクトルを用いて，文の VA 値を回帰予測する．BERT モデルでは，文の先頭に [CLS] トークンを追加し，このトークンが文全体の文脈情報を集約する役割を果たす．[CLS] トークンの 768 次元の特徴ベクトルは，文の感情的ニュアンスを反映しており，それを回帰モデルに入力することで VA 値を予測する．この方法は，文全体の感情情報を効率的に活用できると期待されている．

3.4 方法 2: Attention を活用した予測

方法 2 では，文中の各単語の特徴ベクトルと最終層の Attention 重み [10] を利用して文の VA 値を予測する．具体的には，BERT モデルの最終層で計算された [CLS] トークンへの Attention 重みを使用し，各単語の特徴ベクトルに基づく予測値に重み付けを行う．この重みは，文脈内での単語の重要性を反映しており，重要な単語が VA 値に与える影響を強調する役割を持つ．この方法は，単語間の関連性を考慮しながら文全体の感情情報を統合するため，文脈に基づいた精度の高い VA 値予測が可能であると期待されている．

3.5 評価指標

本研究では，以下の評価指標を使用して回帰モデルと各手法の性能を評価する．

平均二乗誤差（MSE） 回帰モデルの予測精度を測る指標として，予測値と実際の VA 値の差の二乗平均を計算する．MSE は次の式で表される：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

ここで，y_i は実際の VA 値，$\hat{y}_i$ は予測された VA 値であり，MSE が小さいほど予測精度が高いことを示す．

相関係数（Correlation）　予測された VA 値と実際の VA 値の線形的な関係を測定し，感情予測の一致度を示す指標である．相関係数は次の式で計算される：

$$\text{Correlation} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}$$

ここで，$\bar{y}$ および $\bar{\hat{y}}$ は，それぞれ実際の VA 値と予測 VA 値の平均であり，相関係数が 1 に近いほど高い予測精度を示す．

決定係数（R^2）　モデルがデータの変動をどの程度説明できるかを示す指標である．R^2 は次の式で計算される：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

ここで，y_i は実際の VA 値，$\hat{y}_i$ は予測された VA 値，そして $\bar{y}$ は実際の VA 値の平均である．R^2 が 1 に近いほど，モデルの予測精度が高いことを示す．

これらの指標を用いて，各提案手法の予測精度を評価し，最も適切な感情分布モデルを選定する．

4　実験

実験の目的として，提案手法を使用して文の感情価および覚醒度を予測し，それを Emobank データセットの実際の評価値と比較して精度を求める．それにより，各手法に対する精度の違いや，可視化結果の違いを検証する．実験は，三つの主要な部分に分けて実施する．最初に，XANEW データセットを用いて単語レベルの VA 値回帰予測モデルを構築する．次に，二つの手法を用いて文レベルでの VA 値予測を行う．最後に，Russell 円環モデルを用いて感情分布の可視化を行い，Emobank データセットを用いて予測結果を評価する．

4.1　単語レベルの回帰予測モデルの構築

XANEW データセットを基に，BERT を用いて単語ごとの VA 値を回帰予測するモデルを構築した．

実験設定　XANEW データセットの既存の分割（訓練 11,463 件，検証 1,296 件，テスト 1,032 件）を使用した．モデルには BERT による 768 次元の単語埋め込みベクトルを入力し，2 層の隠れ層を持つニューラルネットワークで VA 値を回帰予測する．ReLU 活性化関数と Dropout 層を適用し，Adam Optimizer[11] と早期停止を用いて最適化した．

結果と評価　訓練されたモデルの総合的な MSE は 0.407，R^2 は 0.628 を記録し，全体として良好な性能を示した．また，図 3 に示すように，実際の VA 値と予測値の間に高い相関が確認された．

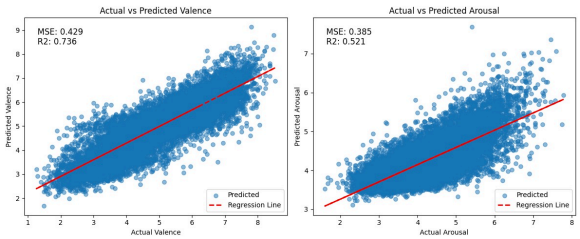

図 3　実際の VA 値と予測された VA 値の散布図

4.2　文の VA 予測値に基づく感情分布

文の準備　GoEmotions データセット（58,009 件）と Kaggle データセット（20,000 件）を用意した．両データセットには異なる感情ラベルが付与されており，それぞれの文を単語に分割し，提案モデルで単語ごとの VA 値を予測した．

感情分布の可視化　文の VA 予測値を感情ラベルごとに平均化し，Russell 円環モデル上で分布を可視化した（図 4〜図 9）．これにより，異なるデータセット間で感情分布の特徴を比較した．

4.3　Emobank データセットによる検証

EmoBank データセットを用いて，各方法で得られた VA 値の精度を MSE と R^2 スコアで評価した．EmoBank は心理学の VAD モデル（Valence，Arousal，Dominance）に基づき，文単位で 1〜5 の範囲で感情値が付与されたデータセットであり，感情分析の分野で広く使用されるベンチマークとして知られている．一方，XANEW データセットの VA 値範囲は 1〜9 であるため，回帰予測モデルが出力する VA 値を EmoBank のスケールに一致させるために線形変換を適用した（例：$V_{Emobank_{predicted}} = 1 + \frac{V_{model}-1}{8} \times 4$）．この正規化により，モデル出力と EmoBank の値を直接比較することが可能となった．

表 1　Emobank データセットにおける各方法の評価結果

手法	Valence		Arousal	
	MSE	Correlation	MSE	Correlation
方法 1	0.917	0.464	0.345	0.180
方法 2	0.311	0.549	0.142	0.257

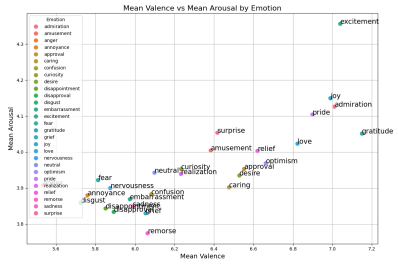

図 4　方法 1：GoEmotions 各感情平均値分布

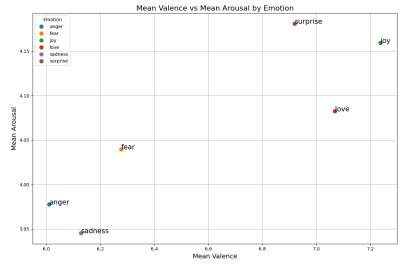

図 5　方法 1：Kaggle データセット各感情平均値分布

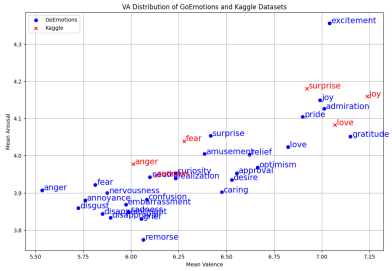

図 6　方法 1：GoEmotions と Kaggle データセット各感情平均値分布

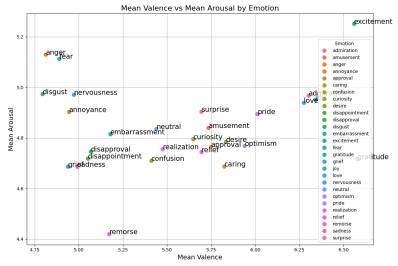

図 7　方法 2：GoEmotions 各感情平均値分布

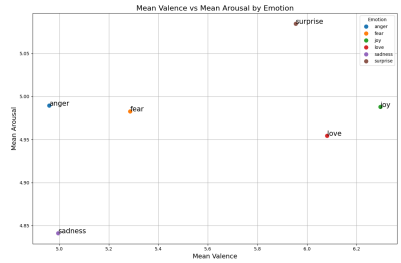

図 8　方法 2：Kaggle データセット各感情平均値分布

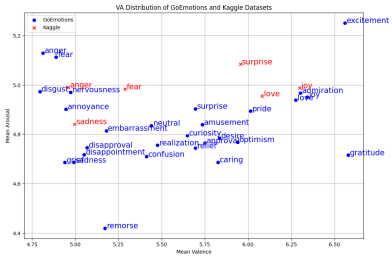

図 9　方法 2：GoEmotions と Kaggle データセット各感情平均値分布

4.4　考察

図 4〜図 9 は，各手法による GoEmotions および Kaggle データセットでの感情平均値分布を示している．また，表 1 の結果によると，方法 2 は方法 1 に比べて MSE がより小さく，相関係数がより高い値を示し，より優れた予測性能を示した．以下にこれらの結果を加え，Russell の円環モデルとの比較を通じて考察する．

図 4〜図 9 において，GoEmotions と Kaggle の感情分布はおおむね類似した傾向を示しているが，いくつかの感情において両データセット間で差異が観察された．そしてポジティブな感情（例：“joy”, “excitement”）は高 Valence 領域，ネガティブな感情（例：“anger”, “sadness”）は低 Valence 領域に分布しており，これは感情の性質に合致している．また，相近の感情が近い位置に分布する傾向も確認された．

方法 1 の結果（図 4〜図 6）は，Valence と Arousal 軸に沿った線状の偏りが見られ，多くの感情が中心付近に集約し，感情間の差異が曖昧である．そして “anger” や “disgust” は，円環モデルの理論的な位置とは異なり，低い Valence の位置に分布している．この結果は，CLS トークンが感情の微細な違いを十分に捉えられていない可能性があることを示唆している．

方法 2（図 7〜図 9）では，感情が多様な領域に広がっており，円環状ではなく，一部の感情は特定の領域に集中する傾向を示している．例えば，“excitement” や “gratitude” などは極端な位置に分布しており，他の感情は中心付近に集まるため，円環モデルのように均等に広がっていない．

5　おわりに

本研究では，XANEW データセットを基にした BERT ベースの回帰モデルを用いて感情分布の解析を行った．提案した 2 つの手法を比較した結果，方法 2 が Emobank データとの一致度が高く、感情間のニュアンスをより正確に捉えたことが確認された（表 1）．この結果から，BERT の Attention 重みを利用した特徴抽出が感情分析に有効であることが示唆された．特徴抽出手法が感情分布の再現性や精度に与える影響が明確になった．

一方，どちらの方法でも感情分布は Russell の円環モデルに見られる均等で対称的な配置とは異なり，偏りや不均一性が見られた．

今後の研究では，さらに高精度な手法や，楕円や非対称的な感情分布を考慮したモデルの構築が求められる．また，Valence や Arousal に加え，Dominance などの新たな次元を追加したモデルの開発が期待される．

参考文献

[1] James Russell. A circumplex model of affect. **Journal of Personality and Social Psychology**, Vol. 39, pp. 1161–1178, 12 1980.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Proceedings of the 31st International Conference on Neural Information Processing Systems**, NIPS'17, p. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[3] A.B. Warriner, V. Kuperman, and M. Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. **Behavior Research**, Vol. 45, No. 4, pp. 1191–1207, 2013.

[4] Sven Buechel and Udo Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 578–585, Valencia, Spain, April 2017. Association for Computational Linguistics.

[5] Marko Horvat, Gordan Gledec, and Fran Leontić. Hybrid natural language processing model for sentiment analysis during natural crisis. **Electronics**, Vol. 13, No. 10, 2024.

[6] Manabu Ito and Konstantin Markov. Sentence embedding based emotion recognition from text data. In **Proceedings of the Conference on Research in Adaptive and Convergent Systems**, RACS '22, p. 53–57, New York, NY, USA, 2022. Association for Computing Machinery.

[7] Margaret M. Bradley and Peter J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. 1999.

[8] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 4040–4054, Online, July 2020. Association for Computational Linguistics.

[9] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: Contextualized affect representations for emotion recognition. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**, pp. 3687–3697, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[10] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupa la, Yonatan Belinkov, and Dieuwke Hupkes, editors, **Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

本附録では，全てのテキストの Valence（感情価）と Arousal（覚醒度）予測結果を示す．それぞれの図は，感情ごとのテキスト予測結果の分布と，その重心位置（赤い十字）を表示する．方法 1 と方法 2 の結果を比較することで，異なる特徴抽出手法が感情分布に与える影響を観察できる．

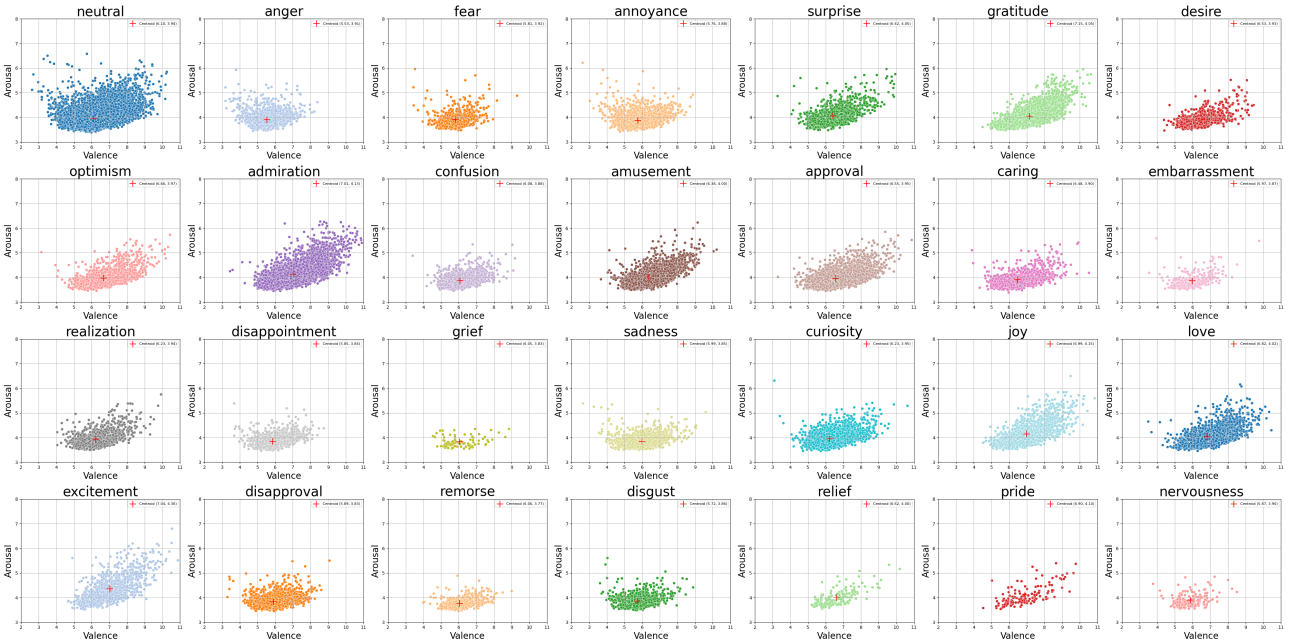

図 10　方法 1 に基づく各感情ラベルのテキスト予測結果における分布

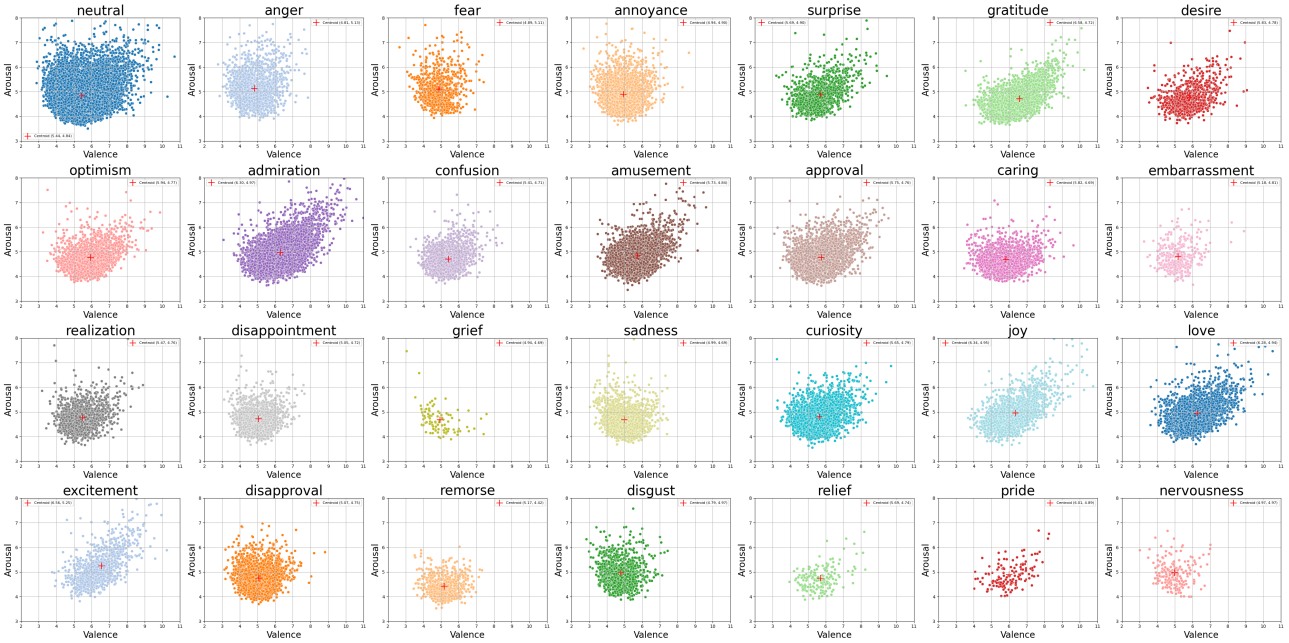

図 11　方法 2 に基づく各感情ラベルのテキスト予測結果における分布