否定語の影響と単語の重要度を考慮した近似 VAD スコアによる感情認識チャットシステムの開発

徐恃源 [1]　尾関智子 [1]
[1] 東海大学大学院
{3meim001,tozeki}@tokai.ac.jp

概要

自然言語処理分野において，データラベリングに高いコストが伴う問題を解決するために，データ拡張および自己教師あり学習手法が大きく発展してきた．感情認識・分類タスクにおいても Park らは，Valence-Arousal-Dominance（VAD）フレームワークを利用することにより，感情のカテゴリラベル付きコーパスを相互に比較・統合し，データセットの拡充ができる可能性があることが示している [1].

本研究では，限られたデータセットを効果的に活用し，感情認識対応チャットシステムを訓練するための新しいアプローチを提案する．具体的には，否定語に影響を受ける動詞と形容詞の V スコアを反転し，LLM のアテンションロジットを加重値として用いることで，感情ラベルが付いていないデータに対して近似 VAD スコアを求める．単に文章中の単語の VAD を平均する従来手法と比べ，精度が向上することを示す.

1　はじめに

Large Language Models (LLMs) のトレーニングおよびチューニングには，膨大な量のデータが必要とされる．LLMs の事前学習においては，自己教師あり学習の適用によりラベル付きデータへの依存は低減しているが，ダウンストリームタスクにおいては相変わらず多量のラベル付きデータが求められる．ラベル付きデータの作成には高いコストが伴うため [2]，データ拡張などを通じたコスト削減を目指す研究が活発に行われている．

感情認識や感情分類のタスクでは，Valence-Arousal-Dominance (VAD)[3] を活用し，Vishnubhotla の研究 [4] のように，単語の VAD スコアの平均を求めることによる文章の VAD スコアの近似方法が一般的に使われている．さらに Park らの VAD フレームワーク [1] は異なるカテゴリでラベルが付けられた自然言語データセットを比較・統合し，VAD を用いることによって，大量のラベル付きデータを必要とする課題に対処できることが示されている．

本研究は，この知見をもとに，会話データセットに VAD を活用することにより，精度の高い感情認識チャットシステムをトレーニングすることを目標とする．

2　関連研究

2.1　VAD モデル

感情を表現する方法として，基本感情に分類するカテゴリー分類が一般的である．これは Ekman の研究 [5] で広く知られており，喜びや悲しみなどの基本的な感情カテゴリに焦点を当てている．一方で，Russell は感情をより柔軟かつ明確に表現するため，多次元空間上の点として感情を表す VA モデルを提案した [6]．さらに，Mehrabian は Dominance（D）次元の重要性を強調し，Valence-Arousal-Dominance（VAD）モデルを提案した [7]．VAD モデルは感情の連続的な特性をより細やかに捉えるために広く活用されている．

2.2　データセット

本研究では，VAD 感情認識に頻繁に使用されるデータセットを用いる．具体的には，1,034 語の英単語が VAD スコアでラベリングされた ANEW[3] と，より大規模な 13,915 語を含む XANEW[8] を利用する．さらに，チャットシステムの開発に特化したデータセットとして，話者に非対称的に情報を伝達することによる知識の非対称性を調整して作成された Topical-Chat[9] と，人によりラベル付けされた DailyDialog[10] を使用する．評価データとしては，既に VAD でラベリングされた EmoBank[11, 12] を用

いる．

3　提案手法

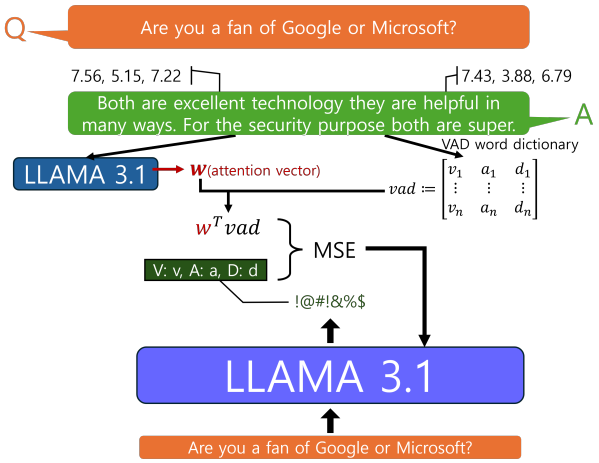

図 1　提案手法

Topical-Chat[9] および DailyDialog[10] から収集した会話データに対して，VAD スコアを推定する必要がある．従来の方法では，文章中の VAD 辞書によりラベル付けされた単語の平均を取ることで文章の近似 VAD スコアを求める手法が広く使われているが，本研究ではより精緻な VAD スコアの近似手法を導入し，モデルの性能向上を図る．

図 1 のように，事前学習済み Llama 3.1-8B モデル[13] をチャットシステムとして使用し，会話データから話者の発話を入力することで生成された応答に対して近似 VAD スコアを推定する．さらに，実際の人間の応答から得られる近似 VAD スコアと，Llama 3.1 モデルが生成した回答文の近似 VAD スコアの間の差を学習の際の平均二乗誤差（MSE）損失関数とすることで，モデルがより人間らしく感情を反映した応答を生成するよう最適化を行う．

テキストデータの前処理および解析の効率化を図るため，広く利用されている自然言語処理ライブラリ spaCy[14] を採用する．spaCy は，Linguistically-motivated トークン化と，固有表現抽出，品詞タグ付け，依存構造解析，文分割，テキスト分類，語彙化（lemmatization），形態素解析など，多様な言語解析コンポーネントを提供し，高速かつ高精度な処理を可能にする．これにより，本研究における否定語の検出および VAD スコア近似の効率と精度を向上させる．

本研究では，ラベルが付いていないデータも感情学習のデータとして用いる方法の検証として，否定語が VAD スコアに与える影響と，アテンションロジットを活用した加重合手法が近似精度に及ぼす改善効果を検証する．

4　実験

以下では，二つの主要な実験を通じて従来手法の限界を評価し，新しい近似手法の有効性を示す．否定語による Valence（V）スコアの誤差や文章中の単語の重要度を考慮し，より正確な VAD スコアの近似手法を検討する．

4.1　否定語の影響

従来の単純な単語の平均を取る方法では，「not」や「never」のような否定語の影響を適切に反映できず，VAD スコアの近似精度が低下するという仮説を立てる．この仮説を検証するために，EmoBank データセットを使用し，（1）否定語を含む文章と含まない文章からなる文章グループ，（2）否定語を含む文章グループ，および（3）否定語を含まない文章グループ，三つのグループを作成し，各文章の本来の VAD スコアと従来手法で求めた近似 VAD スコアを比較する（図 2）．

4.1.1　結果

文章グループ（1）と，従来手法で求めた近似 VAD スコアとの差を基準として，文章グループ（2）および，文章グループ（3）三つのグループの本来の VAD スコアと，従来手法で求めた近似 VAD スコアとの差をボックスプロットで比較する．

文章グループ（2）の場合，文章グループ（1）から約 88%を抽出したグループであるにも関わらず，文章グループ（1）より，V スコアの差は上限が.46，中央値が.03 小さくなったが，A スコアと D スコアの差の場合，ほとんど差がなかった．しかし，文章グループ（3）は，文章グループ（1）から約 12%を抽出したグループでありながら文章グループ（1）と比べ，V スコアの差は上限.01 小さくなったが，中央値が.23 大きくなった．A スコアは上限が.09，中央値が.05 大きくなり，D スコアもまた上限が.15，中央値が.07 大きくなっていた（図 2）．文章グループ（3）と文章グループ（1）の間の A スコアと D スコアの差が発生した原因は文章グループ（2）と文章グループ（1）の A スコアと D スコアの差がほとんどないということから，全体グループである，文章グループ（1）から約 12%だけが抽出されたからであると判断したが，解釈に対して追加的な論議が

必要である可能性がある．

文章グループ（2）および，文章グループ（3）（図 2），文章グループ（3）から否定語の影響を考慮し，否定語が関与する動詞や形容詞の V スコアを反転させた文章グループ（Reversed）（図 3），三つのグループの EmoBank 本来の V スコアとの差の間に顕著な違いが確認される．

特に，否定語の影響を考慮し V スコアを反転させた文章グループ（Reversed）の EmoBank 本来の V スコアとの差と，文章グループ（2）および文章グループ（3）の平均を外れ値の影響を減らすため，ブートストラップで求めた結果，否定語の影響を考慮し V スコアを反転させた文章グループ（Reversed）が文章グループ（3）に比べ，平均の信頼区間が明らかに小さくなっていた（表 1）．

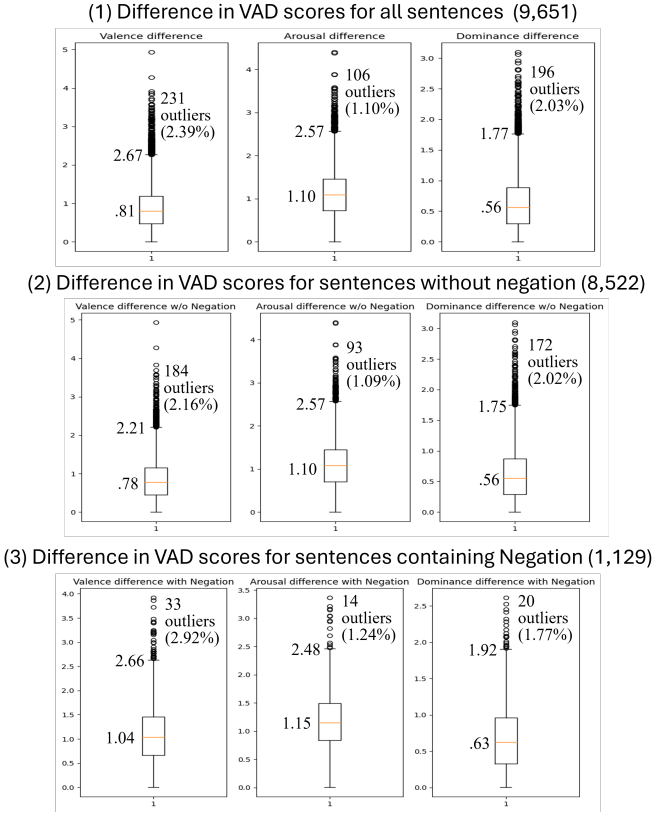

図 2　近似 VAD と文章本来の VAD の間の差

表 1　Bootstrap (2000 resampling) で求めた EmoBank の V スコアと近似 V スコアの間の差の平均

Difference of V	Low	High
(2) W/o negation	.84	.86
(3) With negation	1.08	1.16
Reversed	.89	.96
		α =.05

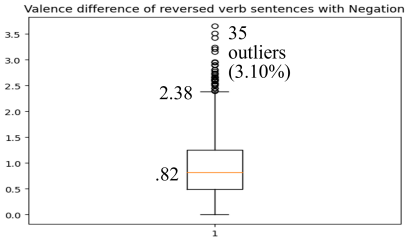

図 3　否定語が入った文章の V スコアの逆転と文章本来の V スコアの差

4.2　加重平均による近似

本研究では，近似 VAD スコアの算出において各単語の重要度を反映するため，事前学習済み LLM の最終アテンションブロックからアテンションロジットを抽出し，各トークンの重みとして用いる．アテンションロジットを重みとして使用するためには，行列を一つのベクトルに変換し，全ての次元を正の値にする必要がある．そのため，以下の六つの手法を二つの複合語の処理の手法に合わせて，12 手法を試す．

(a) Softmax → Sum → L1 Norm
(b) Softmax → Weighted Sum → L1 Norm
(c) Shift_transformation → Sum → L1 Norm
(d) Shift_transformation → Weighted Sum → L1 Norm
(e) Shift_transformation → L1 Norm → Sum → L1 Norm
(f) Shift_transformation → L1 Norm → Weighted Sum → L1 Norm

アテンションロジットの中のアテンションスコアには負数と正数が混在しているため，相対的大きさを維持する L1 正規化を行い，他のヘッダのアテンションロジットと合わせる時，負数のアテンションスコアの影響が極端に減りつつ，正数であることが保証されない問題がある．そのため，負数のアテンションスコアがアテンションロジットにある場合，最小値＋ε（1e-6）を足す Shift transformation を行う．その後，「32 個のアテンションロジットを同じ次元ごとに合算する方法」と，「アテンションヘッドの重要度として解釈することもできる <eos> トークンのアテンションスコアに同じく Shift_transformation と L1 正規化を行う．これを加重値として用い，同じ次元ごとに加重合する方法を検討する．最終的に L1 正規化を行うことにより，加重値として使う．

4.2.1 複合語の処理

複合語とは，複数の語根の組み合わせにより，形成された単語である．Llama 3.1 は，BPE[15] を用いてトークン化を行うため，一つの単語，特に複合語が複数のトークンで分解されることがある．単語の原型に対し，VAD スコアがラベリングされている英単語ディクショナリと一致させる必要があるため，一つの単語が複数のトークンで分解された場合のアテンションスコアを合併させる．合併の方法として，ベクトルの各次元の最大値（max）を求める方法と平均値（mean）を求める方法を考慮する．

4.2.2 結果

近似のために使われる単語の数が少ない場合，正確な評価が出来ないと判断したため，VAD 辞書に含まれている単語（VAD word count）が一定数を超える文章に対し，EmoBank コーパス本来の VAD スコアと上記の方法による近似 VAD スコアの差を求めて平均で表し，文章の中の単語の VAD スコアの単純平均を求める方法を「baseline」，単語の VAD スコアの単純平均を求める方法に否定語の影響（4.1 節）を考慮して求めた VAD スコアの差の平均を「r.baseline」として比較する（表 2）．

複合語の処理に対しては，アテンションスコアを合併させる際, 平均値（mean）より最大値（max）を採る方法がより正確に近似性能が高い．また，<eos> トークンのアテンションスコアを加重値として使い，加重合した場合，単純合をした場合より全般的に精密な近似を示す．以上により，本実験では VAD スコアの近似方法として，「**(f)** Shift_transformation → L1 Norm → Weighted Sum → L1 Norm - max」が適していることがわかる．

表 2 VAD word count が 5 を超える文章の近似 VAD と実際の VAD との差の平均

Difference	V	A	D
baseline	.882	1.117	.582
r. baseline	.866	1.118	.582
(a)-max	.859	1.120	.595
(b)-max	.857	1.093	.595
(c)-max	.858	1.117	.587
(d)-max	.858	1.105	.584
(e)-max	.857	1.116	.588
(f)-max	**.856**	**1.104**	**.586**

5 まとめと今後の課題

本研究ではラベルの付いていない文章に対して VAD スコアをより正確に推定するための方法として二つの方法を文章の中の単語の VAD スコアの単純平均で近似 VAD スコアを求める従来の方法と比較し検証した．V スコアの否定語の影響を考慮するため，否定語が含まれている文章の否定語の影響を受ける動詞や形容詞の V スコアを反転した結果，誤差の信頼区間が約 .2 減少し，否定語が含まれていない文章の近似の誤差に近くなった．Llama 3.1 モデルのアテンションロジットを文章の中の単語の重みとして使う手法として全 12 方法を検証した結果，「Shift_transformation → L1 Norm → Weighted Sum → L1 Norm - max」方法が一番 VAD スコアの近似性能がよく，否定語を反映する方法を適応した結果，VAD 辞書にある単語が五つを超える文章に対して，従来の方法に比べ，近似 V スコアの誤差は減少したが，近似 D スコアの誤差は増加した．これは，文章の中，VAD の近似に用いることができる単語の数が十分である場合，アテンションロジットを適切に加工し，用いることでより正確に VAD スコアの近似ができる可能性があることが示唆されると考えられる．

否定語の影響を反映させるアルゴリズムをより精密化させることで，否定語が影響する単語が名詞である場合の処理を工夫する．さらに，アプローチのモデルの学習の時の損失関数を MSE の代わりに，Park らの VAD フレームワーク [1] で提案された方法をカテゴリ感情でラベル付けさられているデータセット Emotion Detection[16] を用いて試し，本研究で提案したアプローチと並行して用いる方法を考慮する．

参考文献

[1] Sungjoon Park, Jiseon Kim, Seonghyeon Ye, Jaeyeol Jeon, Hee Young Park, and Alice Oh. Dimensional emotion detection from categorical emotion. Association for Computational Linguistics, 2021.

[2] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2020.

[3] Margaret M. Bradley and Peter J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. In **Technical Report C-1**. The Center for Research in Psychophysiology, University of Florida., 1999.

[4] Krishnapriya Vishnubhotla and Saif M. Mohammad. Tweet Emotion Dynamics: Emotion word usage in tweets from US and Canada. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 4162–4176, Marseille, France, June 2022. European Language Resources Association.

[5] Paul Ekman. An argument for basic emotions. **Cognition and Emotion**, Vol. 6, No. 3-4, pp. 169–200, 1992.

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

[7] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. **Current Psychology**, Vol. 14, pp. 261–292, 1996.

[8] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. **Behavior research methods**, Vol. 45, pp. 1191–1207, 2013.

[9] Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In **Proc. Interspeech 2019**, pp. 1891–1895, 2019.

[10] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A manually labelled multi-turn dialogue dataset. In Greg Kondrak and Taro Watanabe, editors, **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 986–995, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[11] Sven Buechel and Udo Hahn. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. Association for Computational Linguistics, April 2017.

[12] Sven Buechel and Udo Hahn. Readers vs. writers vs. texts: Coping with different perspectives of text understanding in emotion annotation. Association for Computational Linguistics, April 2017.

[13] Aaron Grattafiori, et al. The llama 3 herd of models, 2024.

[14] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

[15] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units, 2016.

[16] Sayyed M. Zahiri and Jinho D. Choi. Emotion detection on tv show transcripts with sequence-based convolutional neural networks, 2017.