タスクベクトル演算を用いた感情表現テキスト生成モデル合成手法

天野椋太[1] 目良和也[1] 黒澤義明[1] 竹澤寿幸[1]
[1]広島市立大学大学院 情報科学研究科
{mera, kurosawa, takezawa}@hiroshima-cu.ac.jp

概要

近年の大規模言語モデル（LLM）を活用した対話システムの飛躍的な発展に伴い，対話システムのパーソナライズへの関心も高まってきている．しかし新たな性格のテキスト生成モデルを表現するには，表現したい感情の強さや混ざり具合に応じたモデルを個別に作らなければならない．本研究では，個別の感情を表現することに特化したモデルを感情ごとに作成し，タスクベクトル演算を用いてモデルを合成することで，感情の強弱や複合感情を表現する手法を提案する．LLM を用いたテキスト評価実験の結果，対象感情の強弱と感情の複合いずれについても生成テキストで表現できていることが確認された．

1 はじめに

2022 年 11 月の ChatGPT リリース以降，大規模言語モデルの急速な発展に伴って対話システムの社会実装への関心が高まっている．対話システムはカスタマーサポート，カウンセリング，介護現場など，様々な場面で必要とされているが，近年ではメタバースにおける AI キャラクターやヒトデジタルツインなど，時には対話システム自身が個性を持った応答を行うことが期待される分野もある．例えば，自治体での案内業務を行う対話システムに対して広報キャラのような既存のキャラクターの特徴を持たせるといった研究が行われている[1].

そこで，対話システムに記憶，経歴，性格などの個性を持たせるため，個性を構成する要素を文章や埋め込みベクトルで入力する手法が提案されている[2][3]．しかし個性を文章で表現する手法では表現の微細な調整が困難であることや，個性を詳細に表現するために多量の文章を準備する必要があるといった問題がある．一方，個性を埋め込みベクトルで表現する手法は fine-tuning により実現される．しかし，fine-tuning では新たな性格を表現したいとき，その都度再学習が必要になるという問題がある．

表現したい性格のモデルを個別に作らなければいけないという問題点に対して，個々の観点に沿った応答スタイルのモデルを作成し，それぞれのモデルを合成することで複数の観点を同時に表現可能なモデルを作成する先行研究[4]がある．この先行研究の応用として，応答スタイルの代わりに典型的な性格を表現可能なモデルを作成し，それらのモデルを合成することで再学習を行うことなく複数の性格特徴を併せ持つモデルを構築可能となることが期待できる．ただし，性格を表現するモデルを構築するには，性格がラベル付けされた学習データの収集や，出力テキストの評価が困難である．

そこで本研究では，データ収集や評価が難しい性格の代わりに感情を対象として，**“個々の感情を表現するモデルを合成することで感情の強弱や複合感情を表現できるモデルを生成する”**手法を提案する．また，定量的評価が難しい感情を考慮した応答テキストを評価するため，LLM-as-a-Judge[5]を用いた評価手法も提案する．

2 関連研究

近年，複数のモデル間でモデルパラメータの線形補間や加重平均を行うこと（モデルマージ）で，モデルの能力を操作できることが報告されている．中でも Ilharco ら[6]は，学習前後のモデルパラメータの差分からなるタスクベクトルという概念を定義し，このタスクベクトルを加減算することでモデルの能力を足し引きする手法を提案している．

このタスクベクトルは近年のモデルマージ分野で重要な概念となっている．また，Huang ら[7]では既存の基盤モデルと指示チューニングモデルの重みの差を表す Chat Vector という概念を定義し，図 1 に示すように異なる言語で追加学習されたモデルに対して更なる追加学習なしで指示応答性能を付与する手法を提案している．

また，Jang ら[4]は専門性・情報量・応答スタイルという 3 つの観点別に特化した複数のモデルを学習

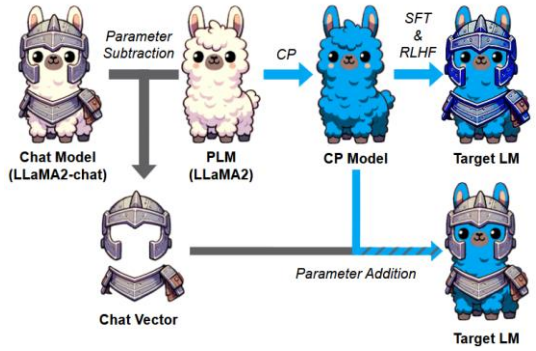

図 1　Chat Vector を用いた性能の付与手法[7]

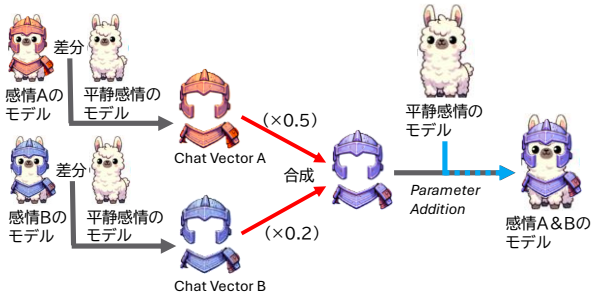

図 2　提案手法による複合性能の付与手法

し，個人の嗜好に合わせてモデルマージを行うことでアライメントをパーソナライズする手法を提案している．Jang らが提案した手法は各モデルに対する係数の和を 1 とする制約付きで，モデルパラメータの加重和によりモデルマージを行うというものである．また，各係数をどのように決定するかに関しては述べられていない．

3 提案手法

3.1 モデル合成による複合感情の表現手法

本研究では，「感情 A のタスクベクトルと感情 B のタスクベクトルを足し合わせたベクトルを適用することで，感情 A と感情 B を併せ持ったテキストを生成するモデルが構築できる」という仮説のもと，図 2 に示すモデル合成手法を提案する．本手法では，平静感情モデルと感情 A モデルの差分から感情 A の Chat Vector を取得し，各感情の Chat Vector を重み付きで合成することで出来た Chat Vector を平静感情モデルに適用することで，複数の感情を併せ持ったモデルを構築する．

ベクトルを合成する際，Chat Vector に対して 1.0 の倍率をかけ合わせてマージすると特化モデルそのものの効果が現れるが，1.0 未満の倍率をかけ合わせてマージすると，特化モデルの影響を軽減すると想定される．また，合成相手のモデルを平静感情モデルにすることで，混ぜられた感情特化モデルは感情を弱く含むテキストを生成するモデルになる．

本手法ではまず平静感情のテキストデータを用いて既存の指示チューニング済みモデルを微調整し，平静感情を表現するモデルを構築する，次に，対象とする感情ごとに平静感情モデルを微調整し，個々の感情の表現に特化したモデルを用意する．そして平静感情モデルと各特化感情モデルの差分から各特化感情のタスクベクトルを算出する．

3.2 データセット

感情を含んだテキスト生成のために，データセットとして WRIME[8]を用いる．WRIME は 80 名の参加者から SNS における投稿計 43,200 件を収集したデータセットであり，投稿者自身による“投稿に含まれる感情”および読み手 3 名による“投稿を読んだ時の感情”の 4 段階評価が付与されている．評価感情は Plutchik の 8 感情モデル[9]に基づく喜び・悲しみ・期待・驚き・怒り・恐れ・嫌悪・信頼の 8 種類であり，各感情について無・弱・中・強の 4 段階で評価が行われている．

本研究では学習に割り当てるデータ数と感情が複合している状態における判断のしやすさを考慮し，喜び・悲しみ・驚き・恐れの 4 感情を使用する．また，投稿者自身の感情に比べて読み手の感情のほうが推定しやすいという報告[8]があることから，本研究では読み手の感情を使って実験を行う．WRIME では読み手の人数分の評価が付与されているため，本研究では読み手の評価の平均を用いる．

投稿には複数の感情が含まれることがあり，すべて使用すると個別感情モデルの学習を妨げる要因となる．そこで，複数の感情が比較的混ざっていない投稿（単独感情投稿）のみを用いる．分類方法を以下に示す．

1. 全ての感情に対して評価値が無 → 平静データ
2. 感情 A の評価値が中 or 強 かつ 感情 A の評価値 ＞感情 A 以外の評価値 → 感情 A の強データ
3. 感情 A の評価値が弱 かつ 感情 A 以外の評価値が無 → 感情 A の弱データ
4. 1〜3 に該当しない投稿データは使用しない

3.3 感情特化モデル学習手法

ベースモデルとして，18 億パラメータの基盤モデ

ル llm-jp-3-1.8b[i]に対して指示チューニングを施したモデルである llm-jp-3-1.8b-instruct[ii]を用いる．まず，ベースモデルに対して平静感情を表現するモデルの学習を行う．次に平静感情モデルを基に，それぞれ異なる感情に特化した 4 つのモデルを学習する．

平静感情モデルの学習は教師あり学習で行う．その際，Low Rank Adaptation（LoRA）[10]を用いて重みを近似する．

各特化感情モデルの学習は Direct Preference Optimization（DPO）[11]によって行う．DPO では学習したい傾向のテキストデータと学習したくない傾向のテキストデータをペアにすることで，モデルの出力が学習したい傾向に近づくような学習を行う．そこで感情の弱データと強データをランダムにペアにする必要があるため，各感情での学習に用いるデータ総数は弱データと強データの少ないほうの件数の 2 倍となる．なお，特化感情の学習時も平静の学習と同様に，LoRA を用いる．

4 評価実験

本評価実験では提案手法が，(1) 単独感情を表現するテキスト生成において設定した感情強度にどの程度従うか，(2) 2 つの感情を表現するテキスト生成タスクにおいて設定した感情強度にどの程度従うか，の 2 点について検証する．

4.1 実験条件

学習時の環境構築ツールとして Transformers[12], trl[iii]，PEFT[iv]を用いる．平静感情学習時は 1epoch，個別感情の学習時は最大で 4epochs 学習する．また，LoRA のパラメータはいずれの場合も rank=8, alpha=16 に設定する．その他の設定は trl のデフォルトに従う．

本実験に用いる単独感情投稿のデータ件数を表 1 に示す．学習データは平静や個別感情の学習に用い，検証データを用いた損失の推移により最適な epoch 数を決定する．

また，ベースラインとして GPT4o-mini[13]を用いた感情表現テキスト生成を行う．その際，対象感情の Train データから異なる強度の投稿をランダムに 3 件取り出し，few-shot プロンプトとして与える．

単独感情の表現実験時の感情の強度は[0.1, 0.2, …, 1.0]の 10 通りとし，各設定で 100 件のテキスト生成を行う．複合感情モデルでの実験時は感情 A の強度 10 通りと感情 B の強度 10 通りの全組み合わせにおいて各 100 件のテキスト生成を行う．単独感情の実験は 4 感情それぞれについて，複合感情の実験は 4 感情のうちの 2 感情の組み合わせ全てについて行う．

表 1　使用した投稿データ件数

	学習データ数	検証データ数
平静	2,401	93
喜び（ペア数）	2,240	83
悲しみ（ペア数）	1,926	54
驚き（ペア数）	1,873	41
恐れ（ペア数）	1,466	27

表 2　人間評価の平均と感情推定器評価の一致率

	喜び	悲しみ	驚き	恐れ
評価者間の平均一致率	0.603	0.393	0.427	0.432
評価者平均と推定器の一致率	0.668	0.539	0.456	0.512

4.2 LLM ベーステキスト感情推定器の構築

テキストに含まれる感情の自動評価のため，LLM を用いてテキストに含まれる感情を推定するシステムを構築した．few-shot プロンプトとして評価例の投稿-スコアのペア 3 件を与え，続けて評価対象の投稿を与えることで，無，弱，中，強の四段階で感情強度を推定させる．使用したモデルは Llama-3.1-70B-Japanese-Instruct-2407[14]である．これは Llama 3.1 70B Instruct[15]に日本語で追加学習を行ったモデルである．軽量化のため，本来は一つあたり 16bit であるパラメータを 4bit に量子化して使用する．

評価手法の信頼性を調べるため，評価者平均と本感情推定手法の評価との一致度を比較した．WRIME の Test データ 2,000 件から few-shot 用の投稿 20 件を確保し，残りの Test データ 1,980 件に対して本評価手法で評価を行った．一致度の指標には Quadratic Weighted Kappa[16]を用いる．一致度の評価結果を表 2 に示す．感情は抽象的なものであるため人間の評価間でも完全に一致するわけではないが，本感情推定器の評価結果は人間同士の評価の一致率と同等以上の一致度を示していることから，本実験では人間による評価の代替として本評価手法を用いる．

[i] https://huggingface.co/llm-jp/llm-jp-3-1.8b
[ii] https://huggingface.co/llm-jp/llm-jp-3-1.8b-instruct
[iii] https://github.com/huggingface/trl
[iv] https://github.com/huggingface/peft

4.3 強度変化による評価の推移

4.3.1 単独感情表現実験

本実験では，モデル合成の際に設定した感情の重みが生成テキストに正しく反映されているかを 4.2 節の感情推定器を用いて確認する．単独感情の強弱についての実験では，対象感情の重みを 0.1 刻みで変えながら，各重み 100 件の生成テキストに対する推定感情強度の比率を算出する．喜び感情での実験結果を図 3 に示す．また，ベースラインである GPT4o-mini の生成テキストにおける感情強度推定結果を図 4 に示す．

提案手法では設定された感情の重みが大きくなるにつれ，生成テキストにおける表出感情の強さもより強くなっていることが確認できる．一方，GPT4o-mini による手法では低レベル領域において正しく感情の強度を変化させられているものの，中レベル以上では強度が飽和していることがわかる．

4.3.2 複合感情表現実験

4.3.1 節と同様の方法で複合感情の強度についても実験を行う．喜びの重みを 0.3 で固定し，驚きの重みを 0.1 刻みで変化させたときの生成テキストに対する驚きの推定強度の分布を図 5 に示す．2 つの感情を出力させる際においても，設定された感情の重みが大きくなるにつれ，より高い推定強度の割合が大きくなっていることが確認できる．

次に，片方の感情の重みを変化させたときもう一方の感情の強度がどれほど影響を受けるか確認する．前述の喜びの重みを固定し驚きの重みを変化させた際の生成テキストに対する喜びの推定強度の推移を図 6 に示す．驚きの重みの変化によって強い喜びの推定割合が多少変化しているが，全体的には一方の感情の重みがもう一方の感情の強度には影響していないことが確認できた．

5 おわりに

本研究では，モデルマージを用いた感情表現テキスト生成手法を提案した．また，喜び・悲しみ・驚き・恐れの 4 感情を用いて提案手法に基づくモデルの構築，および評価実験を行った．その結果，中間的な感情の表現，並びに複数の感情の混合した感情の表現が可能であることを確認した．

今後は，レイヤー単位でマージする手法[17]やタスクベクトル同士の干渉を緩和する手法[18]を取り入れることで，合成モデルのさらなる性能向上，ひいては対話形式のモデルへの応用に取り組みたい．

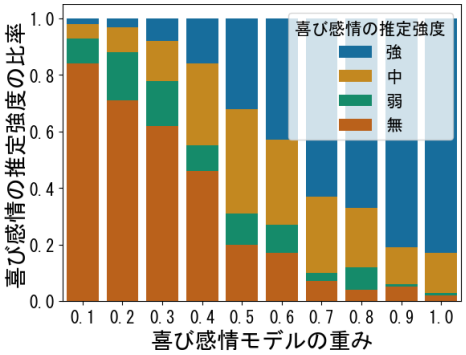

図 3　提案手法による喜び強度の変化

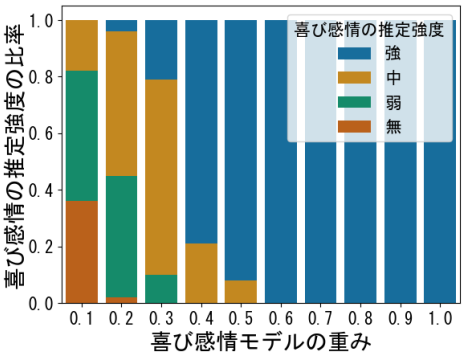

図 4　GPT4o-mini による喜び強度の変化

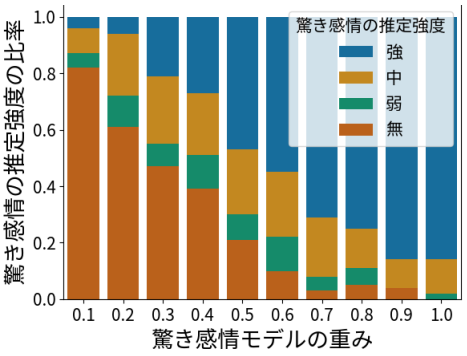

図 5 喜びと驚きの複合モデルにおける驚き強度

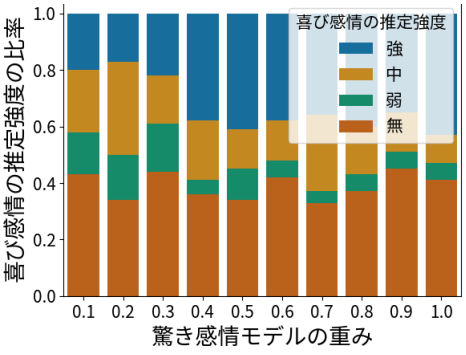

図 6　驚きの重みを変化させた場合の喜び強度

参考文献

1. **水上 雅博, 杉山 弘晃, 有本 庸浩, 東中 竜一郎, 光田 航, 小林 哲生:** “なりきり AI プラットフォームを活用した自治体連携による案内業務対話システム構築”, 人工知能学会論文誌, Vol.38, No.3, pp.B–MA2_1-14. 2023.

2. **S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston:** “Personalizing Dialogue Agents: I have a dog, do you have pets too?”, Proc. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.2204–2213. 2018.

3. **J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan:** “A Persona-Based Neural Conversation Model”, Proc. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.994–1003. 2016.

4. **J. Jang, S. Kim, B. Y. Lin, Y. Wang, J. Hessel, L. Zettlemoyer, H. Hajishirzi, Y. Choi, and P. Ammanabrolu** : “Personalized soups: Personalized large language model alignment via post-hoc parameter merging”, arXiv preprint arXiv:2310.11564. 2023.

5. **L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica:** “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena”, Proc. Advances in Neural Information Processing Systems, Vol.36, pp.46595–46623. 2023.

6. **G. Ilharco, M. T. Ribeiro, M. Wortsman, L. Schmidt, H. Hajishirzi, and A. Farhadi:** “Editing models with task arithmetic”, Proc. The Eleventh International Conference on Learning Representations). 2023.

7. **S. Huang, P. Li, Y. Hsu, K. Chen, Y. T. Lin, S. Hsiao, R. Tsai, and H. Lee:** “Chat Vector: A Simple Approach to Equip LLMs with Instruction Following and Model Alignment in New Languages”, Proc. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.10943–10959. 2024.

8. **T. Kajiwara, C. Chu, N. Takemura, Y. Nakashima, and H. Nagahara:** “WRIME: A New Dataset for Emotional Intensity Estimation with Subjective and Objective Annotations”, Proc. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp.2095–2104. 2021.

9. **R. Plutchik :** “A general psychoevolutionary theory of emotion”, Emotion: Theory, research, and experience, Vol.1. 1980.

10. **E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen:** “LoRA: Low-Rank Adaptation of Large Language Models”, Proc. International Conference on Learning Representations). 2022.

11. **R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn**: “Direct Preference Optimization: Your Language Model is Secretly a Reward Model”, Proc. Advances in Neural Information Processing Systems, Vol.36, pp.53728–53741. 2023.

12. **T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. v. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush:** “Transformers: State-of-the-Art Natural Language Processing”, Proc. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp.38–45. 2020.

13. **OpenAI**. GPT-4o System Card. 2024.

14. **R. Ishigami:** “cyberagent/Llama-3.1-70B-Japanese-Instruct-2407”. 2024.

15. **A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, and A. Fan :** “The llama 3 herd of models”, arXiv preprint arXiv:2407.21783. 2024.

16. **J. Cohen :** “Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit.”, Psychol.Bull., Vol.70, No.4, pp.213. 1968.

17. **T. Akiba, M. Shing, Y. Tang, Q. Sun, and D. Ha :** “Evolutionary optimization of model merging recipes”, arXiv preprint arXiv:2403.13187. 2024.

18. **P. Yadav, D. Tam, L. Choshen, C. A. Raffel, and M. Bansal:** “TIES-Merging: Resolving Interference When Merging Models”, Proc. Advances in Neural Information Processing Systems, Vol.36, pp.7093–7115. 2023.

付録 A　単独感情表現実験で生成されたテキスト

単独感情表現実験の出力テキストと推定器による評価の例を表 A-1 に示す．紙面の都合により一部を省略する．また，参考として WRIME の中から人間の評価者により弱・中・強の評価を受けている投稿を表 A-2 に示す．

表 A-1　提案手法による単独感情の出力事例

設定	提案手法の出力	感情推定結果
喜び 0.2	何もやる気がでない日、一日中寝てられる日なんか、仕事を休むわけにもいかないし、休まれたら迷惑だし、休ませない、て感じだったんかな。そういう状況で仕事してたわけか…	喜び＝無
喜び 0.4	あーもうお腹空いたー(笑)¥n 今日は何作ろうかなぁ(笑)¥n とりあえず昨日作っといたコロッケは冷凍しといて、今日はドリアにしようかなと(笑)	喜び＝中
喜び 0.8	「私はこれで年収 300 万だけど、周りに 1000 万稼ぐ人見たら、なんであんな人と私は違うんだろうって思ったけど、よく考えたら、私の方が楽しいこと大好きで、毎日が文化祭みたいな気分だから、年収 300 万でもいいやって思えた。」って言われて、めっちゃテンション上がった。自分が楽しいって思ったことは、これからも大事にしようと思った。(略)	喜び＝強

表 A-2　WRIME に含まれる投稿

人間が書いた投稿	人間による評価
ちょっと、コーヒーで小休憩	喜び＝弱
人に優しくすると気持ちが良い！	喜び＝中
３１日休みになった??！！！うれしい！！！	喜び＝強

付録 B　複合感情表現実験で生成されたテキスト

複合感情表現実験の出力テキストと推定器による評価の例を表 B-1 に示す．

表 B-1　提案手法による複合感情の出力事例

設定	提案手法の出力	感情推定結果
悲しみ 0.5 ＋ 喜び 0.5	この２週間ほどほぼ毎日飲んでたビールとワインが全く受け付けられなくなり、ほぼ 1 日中お腹が痛いという状態になってしまった orz なんかストレスとかそういうのからきてるらしいんだけど胃の動きが悪くなってるだけだと思ってたから、本当に驚いたし、悲しい気持ちになったわ。。。	悲しみ＝強, 喜び＝無
悲しみ 0.5 ＋ 恐れ 0.5	昨日から胃が痛くて今日はほぼ一日胃痛が収まらない…。胃が痛くてもちゃんと食べないといけないのつらい。普通にご飯食べられないなんて。せっかく久しぶりの友達とごはんの予定だったのに。	悲しみ＝強, 恐れ＝中
悲しみ 0.5 ＋ 驚き 0.5	昨日のテレビでやってた「スマホ老眼」になる人の特徴で、「毎日見るから好き！かっこいい！とか言う人」って紹介されててびっくりした。私、毎日のように好きな人のツイッターのタイムラインを見てしまうんだけど、それって老眼になる前兆なの？？？？？？	悲しみ＝無, 恐れ＝強