International Journal of Research in Interdisciplinary Studies
Volume 4, Issue 11, November 2026
www.ijris.com | E-ISSN: 2584-1017

Affective Dynamics in Text Carry Complementary but Modest Signal for Big Five Personality Prediction: A Rigorous Evaluation on the Essays Corpus

Md Selim Mia[1*], Mishal Naser Al[2]
[1]Department of Artificial Intelligence, Nanjing University of Information Science and Technology (NUIST), Nanjing, China
[2]Department of Computer Science & Technology, Nanjing University of Information Science and Technology (NUIST), Nanjing, China

Abstract—Psychological theory links personality to how emotions unfold over time, yet text-based personality detection usually treats affect as a static summary. We segment each essay of the Essays corpus (n = 2,467) into 64-word windows, score every window with pretrained sentiment and emotion models, and derive 79 trajectory features. Under 3 × 10-fold cross-validation with corrected resampled t-tests, affect dynamics differ significantly between high and low trait groups: valence rises more steeply for high Conscientiousness, Extraversion and Agreeableness, and is lower but less volatile for high Neuroticism. Affect is largely not recoverable from MPNet embeddings (mean cross-validated R^2 = 0.35 for means and 0.11 for dynamics), yet stacking it with embeddings adds only 0.65 macro-F1 points (p = 0.14). The fused model (58.84% accuracy) matches a fine-tuned RoBERTa (58.91%) without GPU training. Affective dynamics are therefore complementary but modest predictors.

***Index Terms*—affective dynamics, Big Five, emotion trajectories, Personality detection, sentiment analysis, transformers.**

1. Introduction

The Big Five traits of Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism (Goldberg, 1990; McCrae and John, 1992) shape how people think, feel and communicate, which makes language a natural source for automatic personality detection (Vinciarelli and Mohammadi, 2014; Mehta et al., 2020b). Methods have evolved from psycholinguistic lexicons such as LIWC (Pennebaker and King, 1999; Mairesse et al., 2007; Tausczik and Pennebaker, 2010) to deep networks, pretrained transformers and large language models (Majumder et al., 2017; Mehta et al., 2020a; Yang et al., 2023).

Emotion links personality and language: Extraversion is associated with positive affect and Neuroticism with negative affect (Costa and McCrae, 1980; Watson and Clark, 1992). Psychology increasingly treats affect as a *dynamic* process, and individual differences in its variability and instability carry information beyond mean levels (Eid and Diener, 1999; Fleeson, 2001; Kuppens et al., 2010; Houben et al., 2015). Computational studies, however, mostly summarise emotion at the document level (Mohammad and Kiritchenko, 2013; Li et al., 2025). It remains untested whether the *temporal shape* of affect within a text reveals personality, and whether contextual embeddings already capture it. In addition, gains reported on the Essays benchmark are often small and rarely tested with statistics that account for overlapping cross-validation folds (Nadeau and Bengio, 2003; Bouckaert and Frank, 2004).

We therefore ask three research questions. RQ1: Do affective dynamics differ between authors high and low on each trait? RQ2: Do they improve prediction beyond contextual text representations? RQ3: Is affective information redundant with, or complementary to, those representations? Our contributions are (i) an interpretable representation of 79 affect-trajectory features, (ii) statistical evidence that valence trends, shifts and volatility differ between trait groups, (iii) a redundancy analysis showing that affect is complementary to embeddings but adds only a modest signal, and (iv) a rigorous benchmark of 15 systems, including a fine-tuned RoBERTa and a zero-shot LLM, with corrected significance tests.

2. Literature and Theoretical Underpinning

A. Personality and Affect Dynamics

Affect is central to the Five-Factor Model, especially to Extraversion and Neuroticism (Costa and McCrae, 1980; Watson and Clark, 1992). Beyond average levels, Fleeson (2001) described traits as distributions of momentary states, Eid and Diener (1999) showed that affective variability is reliable and related to personality, and Kuppens et al. (2010) modelled individual differences in affect dynamics. A meta-analysis by Houben et al. (2015) linked higher emotional variability and instability to lower well-being. These studies suggest that *how* affect changes may reveal personality in addition to *how much* is expressed.

*Corresponding author: selim.shpi@gmail.com

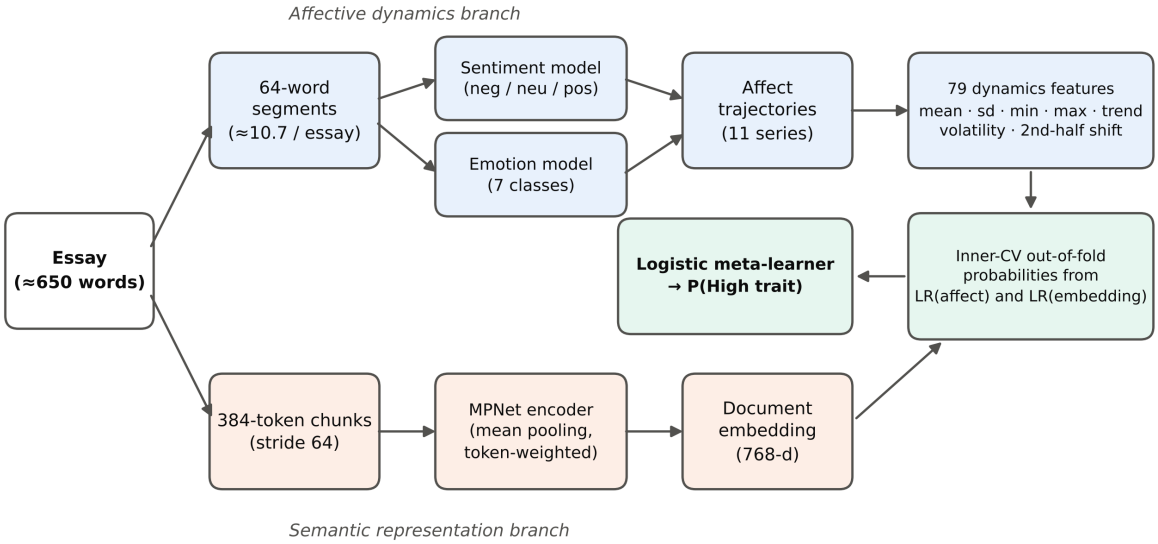


Fig. 1. Overview of the proposed framework

B. Text-Based Personality Detection

Pennebaker and King (1999) released the stream-of-consciousness Essays corpus and showed that linguistic style is a stable individual difference. Mairesse et al. (2007) used psycholinguistic cues for personality recognition, and large-scale studies linked word use on blogs and social media to the Big Five (Yarkoni, 2010; Kosinski et al., 2013; Schwartz et al., 2013). Majumder et al. (2017) combined a convolutional document model with Mairesse features, reaching 56.71–62.68% accuracy per trait on Essays. Transformer-based approaches (Devlin et al., 2019) followed, such as bagged SVMs over BERT embeddings (Kazameini et al., 2020) and hybrids of psycholinguistic and language-model features (Mehta et al., 2020a). Large language models have also been tested as zero-shot recognisers (Ji et al., 2023) and guided by questionnaire-based chains of thought (Yang et al., 2023).

C. Emotion and Within-Text Dynamics

Mohammad and Kiritchenko (2013) showed that fine-grained emotion features improve personality classification on Essays, and EERPD added emotion and emotion-regulation cues to LLM-based detection (Li et al., 2025); both represent emotion at the document level. Kerz et al. (2022) modelled within-text distributions of psycholinguistic features alongside a transformer, and Reagan et al. (2016) showed that sentiment trajectories reveal emotional arcs in stories. Our study differs by (i) focusing on *affective* trajectories summarised with theory-motivated statistics, (ii) testing group differences directly, and (iii) measuring whether affect is redundant with contextual embeddings.

3. Methodology

Figure 1 summarises the approach. Each essay passes through an affective-dynamics branch and a semantic-representation branch, whose predictions are combined by stacked generalisation (Wolpert, 1992).

A. Dataset

We use the Essays corpus (Pennebaker and King, 1999) as distributed by Majumder et al. (2017): 2,467 stream-of-consciousness essays by psychology students, each with a binary (high/low) label for every trait (Table 1). Classes are nearly balanced and inter-trait correlations are weak ($|\varphi| \leq 0.19$).

B. Affect-Dynamics Features

Each essay is split into non-overlapping 64-word windows. Every window is scored by a RoBERTa sentiment model (negative, neutral, positive; Loureiro et al., 2022) and a DistilRoBERTa emotion model covering Ekman's (1992) basic emotions plus neutral (Hartmann, 2022). Together with valence, defined as P(positive) − P(negative), this yields 11 affect series per essay. For each series we compute the seven statistics in Table 2, motivated by the affect-dynamics literature (Eid and Diener, 1999; Kuppens et al., 2010; Houben et al., 2015), and add the polarity shift rate and the mean emotion entropy, giving $7 \times 11 + 2 = 79$ features. The 11 means form the *affect-static* set and all 79 the *affect-dynamic* set.

C. Document Embeddings

Essays are encoded with the all-mpnet-base-v2 encoder,

Table 1
Descriptive statistics of the Essays corpus

Property	Value
Number of essays	2,467
Words per essay, mean (SD)	652 (254)
Words per essay, range	33 – 2,500
64-word segments (mean per essay)	26,353 (10.7)
High class: Openness / Conscientiousness	51.5% / 50.8%
High class: Extraversion/Agreeableness/Neuroticism	51.7% / 53.1% / 50.0%

based on MPNet (Song et al., 2020) and Sentence-BERT (Reimers and Gurevych, 2019). Each essay is split into overlapping 384-token chunks (stride 64), and the mean-pooled chunk vectors are averaged by token count and L2-normalised (768 dimensions).

D. Models

All classifiers are trained per trait. Logistic regression (LR) models standardise their inputs and tune the regularisation strength by inner cross-validation (Pedregosa et al., 2011). The systems are: majority class; TF-IDF (unigrams and bigrams) + LR; affect-static + LR; affect-dynamic + LR; MPNet embedding + LR; early fusion (feature concatenation); probability averaging; and the proposed late fusion, in which inner 5-fold out-of-fold probabilities from the embedding LR and the affect LR train a logistic meta-learner (Wolpert, 1992). Three ablations use affect-static, sentiment-only or emotion-only features.

Two stronger references are included. RoBERTa-base (Liu et al., 2019) is fine-tuned on all five traits using up to four 512-token chunks per essay, with a learning rate of 0.00001, early stopping on a 10% validation split (at most six epochs) and thresholds matched to the training class prior; a variant averages its probabilities with the affect model. A zero-shot LLM, Qwen2.5-3B-Instruct (Qwen Team, 2024), is asked whether the author is High or Low on each trait, and P(High) is read from the next-token logits; we report a 0.5 threshold and a label-free median threshold. Transformer components use Hugging Face Transformers (Wolf et al., 2020).

E. Evaluation Protocol

All systems share the same 3 × 10 repeated cross-validation splits; RoBERTa uses the ten folds of the first repetition. We report accuracy, macro-F1 (primary metric) and ROC-AUC. Differences are tested with the corrected resampled t-test (Nadeau and Bengio, 2003; Bouckaert and Frank, 2004) with Holm (1979) correction across traits. For RQ1 we use point-biserial correlations, Welch t-tests and Cohen's d with Benjamini and Hochberg (1995) FDR control. For RQ3, each affect feature is predicted from the embedding with cross-validated ridge regression (R^2), and the out-of-fold predictions of the affect-only and embedding-only classifiers are correlated. Experiments ran on one NVIDIA T4 GPU with seed 42.

4. Results and Findings

A. Affect Differs Between Trait Groups (RQ1)

The correlations in Figure 2 agree with personality theory. Neuroticism correlates with negative sentiment (r = 0.17),

Table 2
Trajectory statistics for each affect series

Statistic	Definition	Psychological interpretation
Mean	Average over windows	Affect level
SD	Standard deviation over windows	Variability
Min / Max	Extremes over windows	Peak affect
Trend	Slope of a linear fit against position	Direction of change
Volatility	Mean absolute difference between consecutive windows	Instability
2nd-half shift	Mean of second half minus mean of first half	Late-text change

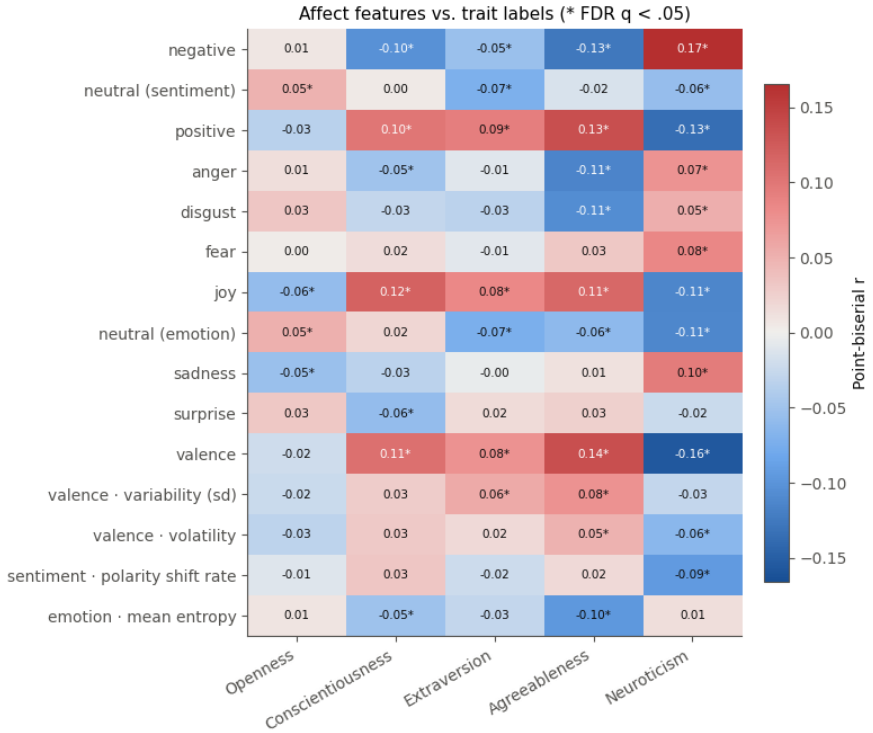


Fig. 2. Affect–trait correlations (* FDR q < .05)

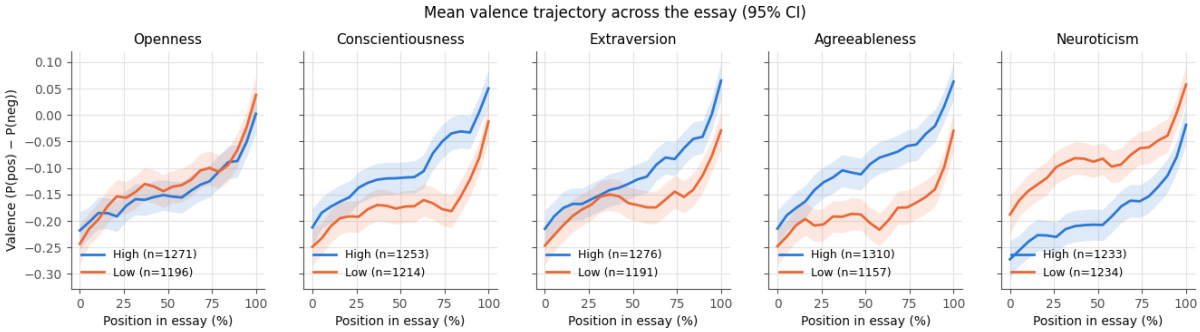


Fig. 3. Mean valence trajectories of high and low trait groups (95% CI)

Table 4
Macro-F1 per trait and average F1, accuracy and AUC (%); † 10 folds

Model	O	C	E	A	N	Avg F1	Acc	AUC
Majority class	33.98	33.65	34.07	34.66	32.29	33.73	50.97	50.00
TF-IDF + LR	62.66	56.10	56.80	53.06	58.33	57.39	58.14	61.54
Affect-static + LR	49.16	52.96	52.72	55.11	58.23	53.64	54.97	57.14
Affect-dynamic + LR	51.28	55.04	53.70	55.77	57.09	54.57	55.42	57.51
MPNet embedding + LR	61.66	56.71	56.33	55.44	58.86	57.80	58.10	61.39
Zero-shot LLM (thr. 0.5)	51.69	35.59	36.84	49.92	47.29	44.27	52.44	58.98
Zero-shot LLM (median thr.)	55.41	55.22	57.66	56.44	58.32	56.61	56.71	58.98
RoBERTa fine-tuned†	63.07	57.03	57.86	58.21	57.31	58.69	58.91	62.02
RoBERTa + affect†	62.61	57.66	56.64	58.11	58.59	58.72	59.13	62.50
MPNet + affect (early fusion)	61.13	58.14	56.71	56.84	59.79	58.52	58.80	62.19
MPNet + affect (prob. avg.)	61.16	58.04	56.33	55.46	60.08	58.21	58.76	62.36
Proposed (late fusion)	61.48	57.82	56.45	56.61	59.90	58.45	58.84	62.43

sadness (0.10) and fear (0.08), and negatively with valence (−0.16) and joy (−0.11). Agreeableness shows the opposite profile, with higher valence (0.14) and joy (0.11) and less anger and disgust (−0.11). Conscientiousness and Extraversion relate positively to joy and valence, while Openness shows almost no affective signature ($|r| \leq 0.06$). As is typical for language–personality associations, all effects are small (Yarkoni, 2010; Schwartz et al., 2013).

Valence rises towards the end of every essay, but more steeply for authors high in Conscientiousness, Extraversion and Agreeableness (Figure 3). Table 3 confirms significantly larger trends (d = 0.10–0.13) and second-half shifts (d = 0.10–0.14) for these traits. High-Neuroticism authors show lower valence (d = −0.32) and, unexpectedly, *less* volatile valence (d = −0.13). Openness shows no differences.

Table 3
Valence dynamics: Cohen's d between high and low trait groups
(* FDR q < .05)

Trait	Mean d	Trend d	Shift d	Volatility d
Openness	−0.04	−0.05	−0.04	−0.06
Conscientiousness	0.21*	0.10*	0.10*	0.06
Extraversion	0.15*	0.12*	0.14*	0.04
Agreeableness	0.28*	0.13*	0.13*	0.10*
Neuroticism	−0.32*	0.03	0.03	−0.13*

B. *Predictive Performance (RQ2)*

Affect features alone predict personality well above chance (54.57% versus 33.73% macro-F1, $p < .001$), and the dynamic set outperforms the static means by 0.93 points (Table 4). Text representations are stronger (TF-IDF 57.39%, MPNet 57.80%). Standard deviations of the average macro-F1 over folds are 1.3–1.9 points.

Adding affect improves every fusion variant. Late fusion raises macro-F1 from 57.80% to 58.45%, accuracy from 58.10% to 58.84% and ROC-AUC from 61.39% to 62.43%, with the largest gains for Agreeableness (+1.17), Conscientiousness (+1.11) and Neuroticism (+1.04) and none for Openness (−0.18) (Figure 4). The meta-learner accordingly gives affect a weight comparable to the embedding for these three traits (e.g., 0.86 versus 0.46 for Agreeableness) but a negative weight for Openness.

Table 5 shows that the proposed model significantly outperforms the majority baseline and the affect-only models. Its gains over the embedding (+0.65, $p = .139$), TF-IDF (+1.06, $p = .098$) and the calibrated LLM (+1.69, $p = .092$) are consistent but not significant, and it is statistically equivalent to the fine-tuned RoBERTa (−0.39, $p = .538$) without GPU training of the classifier. Each partial affect set adds less than the full dynamic set.

Our accuracy (58.84%) equals that of Majumder et al. (2017) and is close to Kazameini et al. (2020) (Table 6), although published results use different splits. EERPD reports 62.40% but was evaluated through LLM APIs on a 10% test subset (Li et al., 2025), so it is not directly comparable.

C. *Redundancy and Complementarity (RQ3)*

The small gains are not caused by redundancy (Figure 5). The embedding recovers affect means only partly (mean $R^2 = 0.35$) and dynamic statistics hardly at all (mean $R^2 = 0.11$; trend and shift $R^2 \leq 0.08$). Predictions of the affect-only and embedding-only classifiers are weakly correlated (r = 0.20–0.42), far less than embedding and TF-IDF (r = 0.48–0.64). Affect dynamics are therefore largely *complementary*, but their unique signal is weak.

D. *Error Analysis*

Only 180 essays (7.3%) are classified correctly on all five

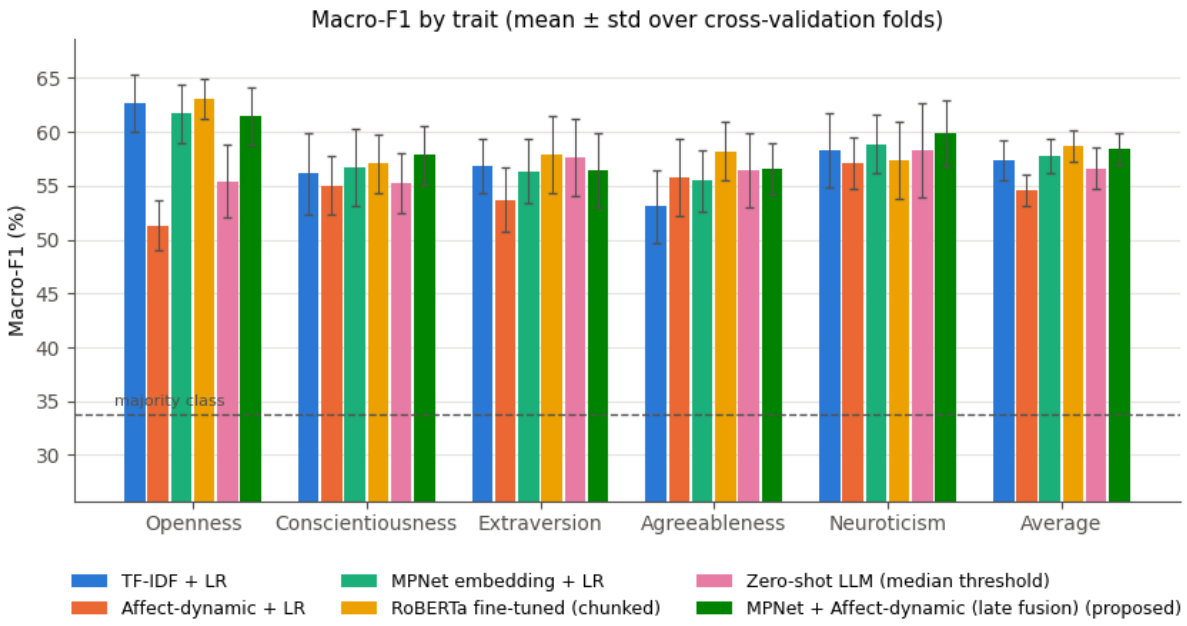


Fig. 4. Macro-F1 by trait for representative systems

Table 5
Proposed model vs. other systems (corrected resampled t-test; † 10 folds)

Comparison (proposed vs.)	Δ Avg F1 (pp)	p	Significant
Majority class	+24.72	< .001	Yes
Affect-static + LR	+4.81	< .001	Yes
Affect-dynamic + LR	+3.88	< .001	Yes
TF-IDF + LR	+1.06	.098	No
MPNet embedding + LR	+0.65	.139	No
Zero-shot LLM (median thr.)	+1.69	.092	No
RoBERTa fine-tuned†	−0.39	.538	No
MPNet + affect-static (late)	+0.25	.333	No
MPNet + sentiment-dynamic only (late)	+0.27	.410	No
MPNet + emotion-dynamic only (late)	+0.40	.110	No
MPNet + affect (early fusion)	−0.07	.819	No

Table 6
Accuracy (%) compared with published results on Essays

System	O	C	E	A	N	Avg
Majumder et al. (2017), CNN + Mairesse	62.68	57.30	58.09	56.71	59.38	58.83
Kazameini et al. (2020), bagged SVM + BERT	62.09	57.84	59.30	56.52	59.39	59.03
RoBERTa fine-tuned (ours)†	63.28	57.24	58.05	58.61	57.40	58.91
MPNet + affect, late fusion (ours)	61.57	57.95	56.89	57.82	59.97	58.84
RoBERTa + affect (ours)†	62.95	57.80	57.15	59.06	58.69	59.13

traits, and accuracy is nearly constant across essay-length quartiles (54.9–61.6%). The most confident errors occur when situational emotion contradicts the label, for example a low-Neuroticism author describing frustration after a computer crash (predicted high with probability 0.84).

5. Discussion

For RQ1, affective dynamics differ between trait groups in theoretically meaningful directions: the higher and rising valence of agreeable, conscientious and extraverted authors fits their positive emotionality (Watson and Clark, 1992), and the lower valence of neurotic authors replicates a classic finding (Costa and McCrae, 1980). Their *lower* volatility contrasts with experience-sampling studies (Eid and Diener, 1999; Kuppens et al., 2010); a likely explanation is that persistently negative text fluctuates less, and a single writing session is a much shorter time scale than days or weeks of self-reports.

For RQ2 and RQ3, affect is a genuine but weak predictor. It is largely invisible to mean-pooled embeddings, which discard order, so fusion helps consistently, yet the gain stays within fold-to-fold noise. Without corrected repeated cross-validation, such a gain could easily be over-reported, and most differences between published Essays systems are of similar size. Two further observations stand out: a calibrated 3B-parameter LLM reaches 56.61% macro-F1 without training, whereas its raw 0.5 threshold collapses (44.27%), and careful fine-tuning makes RoBERTa the strongest single model, with RoBERTa + affect achieving the best accuracy (59.13%) and AUC (62.50%).

Limitations. Essays comes from US undergraduates (1997–2004) with binary median-split labels, the affect models were trained on other domains, only one window size was used, RoBERTa was evaluated on 10 folds and might benefit from longer training, and the LLM may have seen the essays during pre-training.

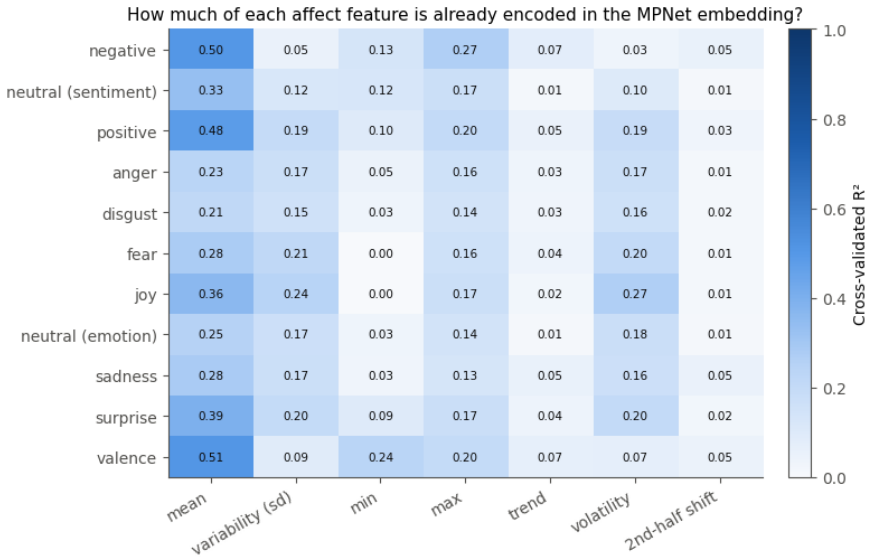


Fig. 5. Cross-validated R^2 of affect features predicted from MPNet embeddings

6. Implication to Research and Practice

Research. Affect dynamics are an interpretable feature family that standard embeddings miss, so models should encode affect order explicitly. Repeated cross-validation with corrected tests should become standard on Essays-style benchmarks, and the redundancy analysis offers a cheap check before fusing new features.

Practice. The late-fusion model matches a fine-tuned transformer while training only logistic regressions, which suits low-resource settings and yields explainable features. However, accuracy near 59% is far from sufficient for individual decisions, and automatic personality inference can enable profiling without consent (Matz et al., 2017). Such systems should not be used for hiring, credit or clinical decisions, and any use requires informed consent and ethical oversight.

7. Conclusion

We examined whether the course of affect within a text reveals Big Five personality. Seventy-nine interpretable affect-dynamics features showed significant, theory-consistent differences between trait groups and proved largely complementary to MPNet embeddings. Fusing them raised macro-F1 from 57.80% to 58.45% and reached 58.84% accuracy, matching a fine-tuned RoBERTa and published state-of-the-art results without transformer training, although the gain was not statistically significant. Affective dynamics are thus a complementary and interpretable, but modest, signal for text-based personality detection.

8. Future Research

Future work should model affect series directly with sequence models, adapt affect classifiers to personal narratives, replicate the analysis on larger corpora with continuous scores, on social media and in other languages such as Bengali, relate text-derived dynamics to experience-sampling data, and combine trajectory evidence with larger LLMs through emotion-aware prompting (Li et al., 2025).

9. Data and Code Availability

The Essays corpus used in this study is publicly available in the repository of Majumder et al. (2017) at https://github.com/SenticNet/personality-detection.

All pretrained models are publicly available on the Hugging Face Hub: cardiffnlp/twitter-roberta-base-sentiment-latest, j-hartmann/emotion-english-distilroberta-base, sentence-transformers/all-mpnet-base-v2, roberta-base and Qwen/Qwen2.5-3B-Instruct. The complete source code is openly available at https://github.com/miamdselim/RIGOROUS-EVALUATION-ON-THE-ESSAYS-CORPUS (notebook: personality-detect-peaper.ipynb). It covers data loading, affect-dynamics feature extraction, document embeddings, all baseline and fusion models, the fine-tuned RoBERTa and zero-shot LLM experiments, statistical tests and figure generation. The notebook runs end-to-end on a single NVIDIA T4 GPU with a fixed random seed (42), so all reported results can be reproduced.

References

[1] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. R. Stat. Soc. Ser. B, Methodol.*, vol. 57, no. 1, pp. 289–300, 1995.

[2] R. R. Bouckaert and E. Frank, "Evaluating the replicability of significance tests for comparing learning algorithms," in *Advances in Knowledge Discovery and Data Mining (PAKDD 2004)* (Lecture Notes in Computer Science), vol. 3056, Berlin, Germany: Springer, 2004, pp. 3–12.

[3] P. T. Costa and R. R. McCrae, "Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people," *J. Pers. Soc. Psychol.*, vol. 38, no. 4, pp. 668–678, Apr. 1980.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol. (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[5] M. Eid and E. Diener, "Intraindividual variability in affect: Reliability, validity, and personality correlates," *J. Pers. Soc. Psychol.*, vol. 76, no. 4, pp. 662–676, Apr. 1999.

[6] P. Ekman, "An argument for basic emotions," *Cogn. Emotion*, vol. 6, no. 3–4, pp. 169–200, May–Jul. 1992.

[7] W. Fleeson, "Toward a structure- and process-integrated view of personality: Traits as density distributions of states," *J. Pers. Soc. Psychol.*, vol. 80, no. 6, pp. 1011–1027, Jun. 2001.

[8] L. R. Goldberg, "An alternative 'description of personality': The Big-Five factor structure," *J. Pers. Soc. Psychol.*, vol. 59, no. 6, pp. 1216–1229, Dec. 1990.

[9] J. Hartmann, "Emotion English DistilRoBERTa-base," Hugging Face, 2022. [Online]. Available: https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

[10] S. Holm, "A simple sequentially rejective multiple test procedure," *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979.

[11] M. Houben, W. Van Den Noortgate, and P. Kuppens, "The relation between short-term emotion dynamics and psychological well-being: A meta-analysis," *Psychol. Bull.*, vol. 141, no. 4, pp. 901–930, Jul. 2015.

[12] Y. Ji, W. Wu, H. Zheng, Y. Hu, X. Chen, and L. He, "Is ChatGPT a good personality recognizer? A preliminary study," arXiv:2307.03952, 2023.

[13] A. Kazameini, S. Fatehi, Y. Mehta, S. Eetemadi, and E. Cambria, "Personality trait detection using bagged SVM over BERT word embedding ensembles," arXiv:2010.01309, 2020.

[14] E. Kerz, Y. Qiao, S. Zanwar, and D. Wiechmann, "Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features," in *Proc. 12th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal. (WASSA)*, Dublin, Ireland, 2022, pp. 182–194.

[15] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 15, pp. 5802–5805, Apr. 2013.

[16] P. Kuppens, Z. Oravecz, and F. Tuerlinckx, "Feelings change: Accounting for individual differences in the temporal dynamics of affect," *J. Pers. Soc. Psychol.*, vol. 99, no. 6, pp. 1042–1060, Dec. 2010.

[17] Z. Li, S. Li, D. Zhu, Q. Ma, and W. Xiong, "EERPD: Leveraging emotion and emotion regulation for improving personality detection," in *Proc. 31st Int. Conf. Comput. Linguistics (COLING 2025)*, Abu Dhabi, UAE, 2025, pp. 7721–7734.

[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv:1907.11692, 2019.

[19] D. Loureiro, F. Barbieri, L. Neves, L. Espinosa Anke, and J. Camacho-Collados, "TimeLMs: Diachronic language models from Twitter," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations*, Dublin, Ireland, 2022, pp. 251–260.

[20] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *J. Artif. Intell. Res.*, vol. 30, pp. 457–500, Nov. 2007.

[21] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 74–79, Mar.–Apr. 2017.

[22] S. C. Matz, M. Kosinski, G. Nave, and D. J. Stillwell, "Psychological targeting as an effective approach to digital mass persuasion," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 48, pp. 12714–12719, Nov. 2017.

[23] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *J. Pers.*, vol. 60, no. 2, pp. 175–215, Jun. 1992.

[24] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi, "Bottom-up and top-down: Predicting personality with psycholinguistic and language model features," in *Proc. 20th IEEE Int. Conf. Data Mining (ICDM)*, Sorrento, Italy, 2020, pp. 1184–1189.

[25] Y. Mehta, N. Majumder, A. Gelbukh, and E. Cambria, "Recent trends in deep learning based personality detection," *Artif. Intell. Rev.*, vol. 53, no. 4, pp. 2313–2339, Apr. 2020.

[26] S. M. Mohammad and S. Kiritchenko, "Using nuances of emotion to identify personality," in *Proc. 7th Int. AAAI Conf. Web Social Media (ICWSM)*, vol. 7, no. 2, 2013, pp. 27–30.

[27] C. Nadeau and Y. Bengio, "Inference for the generalization error," *Mach. Learn.*, vol. 52, no. 3, pp. 239–281, Sep. 2003.

[28] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011.

[29] J. W. Pennebaker and L. A. King, "Linguistic styles: Language use as an individual difference," *J. Pers. Soc. Psychol.*, vol. 77, no. 6, pp. 1296–1312, Dec. 1999.

[30] Qwen Team, "Qwen2.5 technical report," arXiv:2412.15115, 2024.

[31] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds, "The emotional arcs of stories are dominated by six basic shapes," *EPJ Data Sci.*, vol. 5, Art. no. 31, 2016.

[32] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 3982–3992.

[33] H. A. Schwartz *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLOS ONE*, vol. 8, no. 9, Art. no. e73791, Sep. 2013.

[34] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MPNet: Masked and permuted pre-training for language understanding," in *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020, pp. 16857–16867.

[35] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Soc. Psychol.*, vol. 29, no. 1, pp. 24–54, Mar. 2010.

[36] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 273–291, 3rd Quart., 2014.

[37] D. Watson and L. A. Clark, "On traits and temperament: General and specific factors of emotional experience and their relation to the five-factor model," *J. Pers.*, vol. 60, no. 2, pp. 441–476, Jun. 1992.

[38] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process.: Syst. Demonstrations*, 2020, pp. 38–45.

[39] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[40] T. Yang, T. Shi, F. Wan, X. Quan, Q. Wang, B. Wu, and J. Wu, "PsyCoT: Psychological questionnaire as powerful chain-of-thought for personality detection," in *Findings Assoc. Comput. Linguistics: EMNLP 2023*, Singapore, 2023, pp. 3305–3320.

[41] T. Yarkoni, "Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers," *J. Res. Pers.*, vol. 44, no. 3, pp. 363–373, Jun. 2010.