

Article

Learning When to Feel: Scalar-Gated Fusion and Affective Flow Representations

Hiram Calvo ^{1,2,*} , Mayte H. Laureano ^{1,2}, Pablo Gervás ^{2,3}  and Gonzalo Méndez ^{2,3} 

¹ Center for Computing Research, Instituto Politécnico Nacional, Mexico City 07738, Mexico; mhernandezl2021@cic.ipn.mx

² Facultad de Informática, Universidad Complutense de Madrid, 28040 Madrid, Spain; pgervas@ucm.es (P.G.); gmendez@ucm.es (G.M.)

³ Instituto de Tecnología del Conocimiento, Universidad Complutense de Madrid, 28040 Madrid, Spain

* Correspondence: hcalvo@cic.ipn.mx

Abstract

In this paper we study how external affective information should be integrated into a compact transformer-based text classifier. Rather than treating affective features as signals to be appended directly to the representation, we examine whether their contribution should be controlled through lightweight fusion mechanisms. The comparison focuses on scalar-gated fusion versus plain concatenation, using DistilBERT as the textual backbone and four affective resources: the NRC VAD Lexicon, VAD-BERT, Ekman-style emotion scores, and SenticNet. The evaluation is conducted on two English corpora with different label structures: a seven-class MentalHealth dataset and the fine-grained GoEmotions benchmark. Across both corpora, scalar gating consistently matches or outperforms concatenation in terms of Macro-F1. On MentalHealth, scalar gating improves all directly comparable configurations. On GoEmotions, it achieves the best overall Macro-F1 and improves most matched comparisons. Beyond static feature integration, we introduce affective flow (EmoFlow) representations derived from VAD-BERT, which model the evolution of valence, arousal, and dominance across segments of a text. These dynamic representations do not surpass the strongest static lexical resources in absolute performance, but they provide consistent improvements within the VAD-BERT family, particularly when combined with scalar gating or cross-attention. Our contribution is twofold. First, we show that a lightweight scalar gate provides an effective and interpretable mechanism for adaptively integrating low-dimensional affective side information into transformer-based classifiers. Second, we introduce affective flow representations that explicitly model how affect evolves within a document, enabling the analysis of both adaptive resource selection and intra-document affective dynamics. Together, these results suggest that the key issue is not only which affective resources to use, but also when and how they should influence the model.



Academic Editors: Tianyong Hao, Guangyou Zhou, Xinyu He and Lu Xiang

Received: 12 May 2026

Revised: 3 July 2026

Accepted: 8 July 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: affective computing; scalar gating; feature fusion; transformer models; DistilBERT; VAD; VAD-BERT; Ekman emotions; SenticNet; emotion classification; mental health classification; transfer learning; NLP

MSC: 68T07; 68T50

1. Introduction

Transformer encoders already capture a large amount of affective and semantic information, but external emotion resources can still provide useful inductive bias, especially in domains where subtle emotional tone matters for the label space. This is particularly relevant in mental health text classification, where distinctions between classes may depend not only on topic words but also on intensity, polarity, and emotional profile. Earlier work on emotion analysis and affective representation has shown the value of both dimensional and categorical views of emotion [1–3]. More recent studies have further highlighted the complementary role of external affective representations, adaptive fusion strategies, and explainable transformer architectures for sentiment and emotion analysis [4–6]. In parallel, commonsense affective resources such as SenticNet offer concept-level information that goes beyond isolated word polarity [7].

In this study, we address the question of how these resources should be fused with a transformer sentence representation. The standard baseline is concatenation: the text vector and the external features are simply appended and passed to the classifier. This approach is easy to implement and often effective, but it treats all side information as equally relevant across instances. In practice, this assumption is strong. Affective resources differ in granularity and noise profile. Lexical VAD is transparent but sparse, VAD-BERT is contextual but model-dependent, Ekman-style features summarize discrete emotional tendencies, and SenticNet provides concept-level commonsense affect with variable coverage. A fusion mechanism that can decide how much to trust each resource for each example may therefore be preferable, motivating adaptive fusion strategies that dynamically regulate heterogeneous information sources rather than assuming equal relevance across instances [5,8].

For this reason, we compare concatenation with scalar gating. Although gating mechanisms have been widely used in multimodal fusion and representation learning, most existing approaches employ high-dimensional vector gates or modality-specific gating networks designed for rich feature spaces [8–12]. Our objective is different. We investigate whether a deliberately simple scalar gate is sufficient for integrating compact affective side information into a transformer classifier. The proposed gate learns a single coefficient for each affective resource and each input instance, providing an adaptive yet lightweight mechanism that regulates the contribution of external affective information. This design is well suited to low-dimensional affective representations such as lexical VAD, contextual VAD estimates, or emotion-category summaries, where more expressive gating architectures may introduce unnecessary parameters without improving interpretability. Moreover, the scalar formulation allows direct inspection of the learned gate values, making it possible to analyze when the model chooses to rely on affective information.

However, affective information in text is not only heterogeneous across resources; it is also internally structured within each document. A single document-level affective vector may hide local variations in emotional tone that are relevant for classification. For instance, a text may transition from neutral to distressed language or mix conflicting emotional cues. This motivates a second extension explored in this work: *affective flow representations*. Instead of summarizing affect at the document level, we derive segment-level VAD signals using VAD-BERT and model their evolution across the text. This idea is consistent with recent studies showing that temporal affective structure often provides complementary information beyond static affective summaries [6,13]. The resulting representations, referred to as EmoFlow variants in the experimental section, encode both local affective states and their first-order transitions, allowing the model to explicitly capture how emotion unfolds within the input.

This extension naturally complements the gating perspective. While scalar gating regulates when affective information should influence the model, affective flow representations describe how affect evolves within the text. Together, they provide a structured view of affective integration by combining adaptive resource weighting with explicit modeling of intra-document dynamics. To study this interaction, we evaluate several fusion mechanisms for affective flow, including concatenation, scalar gating, and cross-attention, the latter allowing the textual representation to selectively attend to different parts of the affective trajectory [11,14].

Accordingly, this work pursues two complementary objectives. The first is to evaluate whether a lightweight scalar-gated fusion strategy provides an effective and interpretable mechanism for integrating low-dimensional affective side information into transformer-based classifiers. The second is to extend this framework with affective flow representations that model intra-document affective dynamics, allowing us to investigate not only whether affective information is useful, but also how its temporal evolution contributes to classification. Beyond predictive performance, we also analyze the behavior of the proposed fusion mechanisms to better understand when affective information contributes to the final decision.

We use DistilBERT as the textual encoder. This choice is motivated by efficiency and stability considerations. Sanh et al. [15] show that DistilBERT reduces model size by about 40%, is significantly faster, and retains most of the language understanding performance of the original BERT. For repeated multi-seed experiments across multiple fusion strategies, this efficiency–performance balance is particularly advantageous.

Related Work

Recent work has explored the integration of affective signals with transformer-based models for emotion prediction and sentiment analysis. For instance, systems in the SemEval-2025 Task 11 framework combine BERT embeddings with VAD-based numerical features to estimate emotion intensity [16]. Similarly, prior approaches incorporate dimensional or categorical affective features as auxiliary inputs to improve classification performance [17,18]. These lines of work suggest that affective resources can provide useful complementary information to contextual language representations, particularly when the target labels are emotionally nuanced or when the input contains implicit affective cues.

Gated fusion mechanisms have also been widely studied in multimodal settings, where gates regulate the contribution of different input sources [9–11]. More recently, adaptive gated fusion and attention-based architectures have been proposed for explainable sentiment analysis and multimodal emotion recognition, showing the benefits of dynamically weighting heterogeneous information sources [5,8,12]. However, these approaches typically rely on vector-valued gates, modality-specific controllers, or high-capacity fusion networks designed for rich multimodal representations such as text, audio, and visual streams.

In contrast, the present work focuses on affective integration in a low-dimensional setting. Rather than relying on high-capacity vector gates, we adopt a lightweight scalar-gated mechanism that aligns with the compact nature of VAD-based representations. Specifically, we investigate whether a single scalar gate is sufficient to regulate the contribution of low-dimensional affective features, providing an adaptive yet interpretable alternative to more expressive gating architectures. Moreover, we extend beyond static affective features by introducing affective flow representations, which explicitly model intra-document emotional dynamics through segment-level VAD trajectories. While recent work has increasingly explored adaptive fusion and explainable emotion modeling, comparatively little attention has been devoted to combining lightweight scalar gating with dynamic affective representations for transformer-based text classification [6,13].

The contribution of this paper is therefore twofold. First, we provide a controlled comparison between concatenation and scalar-gated fusion for integrating heterogeneous affective resources into a transformer-based classifier. Second, we extend this comparison with affective flow representations derived from VAD-BERT, showing how dynamic affective signals can be incorporated through lightweight fusion mechanisms. The results highlight the importance of adaptive integration and structured affective representations, suggesting that classification performance depends not only on the choice of affective features, but also on how they are incorporated into the model.

The remainder of the paper is organized as follows. Section 2 introduces the corpora and affective resources used in the experiments. Section 3 presents the fusion methods, including concatenation, scalar gating, and the proposed affective flow representations, along with their associated fusion mechanisms such as cross-attention. Section 4 describes the experimental setup and evaluation protocol. Section 5 reports and discusses the results, including both aggregate performance and inferential analysis based on bootstrap testing. Finally, Section 6 concludes the paper and outlines directions for future work, including affective transfer learning and hybrid static–dynamic representations.

2. Corpora and Affective Resources

2.1. MentalHealth Corpus

The MentalHealth corpus used in this work comes from the public *Sentiment Analysis for Mental Health* aggregation released on Kaggle [19]. Later work describes the bundled corpus as a collection of 52,681 English social-media statements annotated with seven labels: *Anxiety*, *Bipolar*, *Depression*, *Normal*, *Personality disorder*, *Stress*, and *Suicidal* [20]. The texts are not clinical notes and the labels should not be interpreted as formal diagnoses; rather, they reflect a benchmark-style categorization of mental health-related social-media statements [20]. In our working split, the corpus is divided into 36,876 training instances, 7902 validation instances, and 7903 test instances.

This corpus is particularly suitable for studying affective fusion because emotional information is clearly relevant but not sufficient on its own. Categories such as anxiety, stress, depression, and suicidal discourse exhibit characteristic affective profiles, yet they also share substantial lexical overlap. As a result, models must integrate contextual semantics with affective cues, rather than relying exclusively on either signal. This makes the dataset a natural testbed for evaluating whether external affective features provide complementary information to transformer-based representations.

2.2. GoEmotions

GoEmotions is a large English corpus introduced by Google Research for fine-grained emotion recognition [21]. It contains approximately 58k comments annotated with 27 emotion categories plus a *Neutral* label. Compared with the MentalHealth corpus, GoEmotions offers a broader and more granular emotional label space, covering subtle distinctions such as *admiration*, *confusion*, or *remorse*.

In the present study, GoEmotions serves two complementary roles. First, it allows us to evaluate the fusion mechanisms in a setting where affect is the primary prediction target, rather than an auxiliary signal. Second, it provides a contrastive benchmark to assess whether trends observed in mental health classification generalize to a different domain and label structure. This dual use is important for determining whether improvements associated with specific fusion strategies reflect general modeling properties or dataset-specific effects.

2.3. Affective Resources

We consider four types of affective resources, covering both dimensional and categorical representations, as well as lexical and contextual sources.

The first resource is the NRC VAD Lexicon, which provides human ratings of valence, arousal, and dominance for more than 20,000 English words [2]. For each document, we compute aggregate statistics over matched lexical items, including mean and dispersion measures. The main advantage of this resource is interpretability: the resulting features directly reflect human-annotated affective scores. However, lexical coverage is inherently limited, and some inputs contain few or no matched entries.

The second resource is VAD-BERT. In this paper, the term refers to the public model Robrokools/vad-bert, a BERT-based regressor that outputs sentence-level valence, arousal, and dominance values [22]. According to the model card, it was fine-tuned on EmoBank [22], a benchmark of approximately 10k English sentences annotated in the VAD space [3]. Unlike lexical VAD, VAD-BERT is contextual: it can infer affect even when no explicit affective words are present. In addition to document-level predictions, we use this model to derive segment-level VAD trajectories, which are later used to construct affective flow (EmoFlow) representations.

The third resource is Ekman-style categorical emotion information. Ekman's framework of basic emotions—including anger, disgust, fear, joy, sadness, and surprise—remains one of the most influential categorical accounts of affect [1]. In computational settings, these categories can be represented as a fixed-length emotion score vector. They provide a complementary perspective to VAD-based representations, as categorical and dimensional views capture different aspects of affect. Prior work has shown that these representations can be mapped onto each other in a principled way [4].

The fourth resource is SenticNet, a concept-level affective commonsense resource [7]. Unlike standard word-level lexicons, SenticNet encodes affective information at the level of concepts, including multiword expressions and semantic relations. This allows it to capture affective meaning that is not reducible to individual tokens. For each document, we extract concept-level affective features based on detected concepts and aggregate them into a compact representation.

Taken together, these resources span a range of design choices: lexical versus contextual signals, dimensional versus categorical representations, and word-level versus concept-level coverage. This diversity is intentional. It creates a challenging fusion setting in which the model must reconcile heterogeneous sources of affective information. In the following section, we describe how these signals are integrated into a transformer-based architecture through concatenation, scalar gating, and affective flow representations.

3. Methods

We study how affective information can be integrated into a transformer-based classifier under two complementary settings: static feature augmentation and dynamic affective representations. The focus is not only on which resources are useful, but on how their contribution should be controlled within the model.

We first define two baseline fusion strategies for combining textual and affective signals (Section 3.1): direct concatenation and scalar-gated fusion. These methods differ in whether auxiliary features are always injected into the representation or modulated on a per-instance basis.

We then describe the construction of affective features from multiple sources (Section 3.2), including lexical, contextual, categorical, and concept-level representations, and how they are normalized prior to fusion.

Next, we introduce affective flow representations derived from VAD-BERT (Section 3.3). These representations extend document-level affect modeling by capturing intra-textual variation through segment-level signals and their transitions, and are combined with the textual backbone using scalar gating and cross-attention mechanisms.

Finally, we specify the classification layer, loss functions, normalization procedures, and training protocol (Section 3.3.3), ensuring a controlled and comparable evaluation across all variants.

3.1. Fusion Methods

We consider two main strategies for integrating textual representations with external affective features: direct concatenation and gated fusion. These approaches differ in how they treat the relevance of auxiliary information. Concatenation assumes that all features should be equally available to the classifier, whereas gating introduces an explicit mechanism to modulate their contribution on a per-instance basis. This contrast allows us to isolate the effect of controlled versus unconditional feature injection.

Unless otherwise stated, all projection matrices and bias terms introduced in the following fusion mechanisms are trainable parameters learned jointly with the classifier through backpropagation. The subscript identifies the role of each parameter (e.g., feature projection, gating, or attention). The textual representation is denoted by $h \in \mathbb{R}^d$, where d is the hidden dimension of the DistilBERT backbone. Projection matrices (W) perform linear transformations that map affective features into the shared hidden space of the textual encoder, bias terms (b) are learnable offsets, α denotes a scalar gate, and g denotes a vector-valued gate. Unless explicitly stated otherwise, all fusion parameters are optimized jointly with the classifier, whereas the VAD-BERT regressor remains frozen throughout training and serves exclusively as a fixed feature extractor.

3.1.1. Concatenation Baseline

Let $h \in \mathbb{R}^d$ be the DistilBERT sentence representation for an input text. Let $z_1, \dots, z_R$ denote the external affective feature vectors extracted from R resources, after feature normalization. In the concatenation baseline, fusion is performed by direct appending:

$$h_{\text{concat}} = [h \parallel z_1 \parallel \dots \parallel z_R]. \quad (1)$$

The classifier receives the whole augmented vector and learns a decision boundary over the joint space. This method is simple and transparent. However, it does not explicitly model whether a given resource is reliable for a particular input.

3.1.2. Scalar-Gated Fusion

Our scalar-gated fusion is deliberately lightweight. Each external vector z_r is first projected into the text hidden space:

$$\tilde{z}_r = W_r z_r + b_r, \quad \tilde{z}_r \in \mathbb{R}^d. \quad (2)$$

A scalar gate is then computed for each resource:

$$\alpha_r = \sigma(w_r^\top [h \parallel \tilde{z}_r] + c_r), \quad \alpha_r \in (0, 1), \quad (3)$$

where σ is the sigmoid function. The final fused representation is:

$$h_{\text{gate}} = h + \sum_{r=1}^R \alpha_r \tilde{z}_r. \quad (4)$$

The classifier is applied to h_{gate} .

This design means that each side resource contributes through a single confidence-like coefficient. If the affective signal appears useful for a given text, the gate approaches 1 and the projected feature is injected more strongly. If it appears unreliable or redundant, the gate moves toward 0 and the model falls back mainly on the text representation.

3.1.3. Rationale for Scalar-Gated Fusion

There are three reasons why this mechanism is plausible. First, it is an instance-wise selector. Affective side information is not uniformly helpful. Some inputs have clear lexical affect, while others are emotionally implicit. Scalar gating allows the model to react to that variation instead of forcing the same fusion pattern everywhere, in line with the broader motivation of gated multimodal fusion [10].

Second, it is parameter-efficient. A vector gate learns one coefficient per hidden dimension, which is more flexible but also easier to overfit when side information is low-dimensional. In our experiments, the vectorial VAD gating variant is weaker than scalar VAD gating on MentalHealth. This supports the idea that a single resource-level decision is enough for compact affective signals.

Third, scalar gating preserves the textual backbone as the default carrier of meaning. The residual form in Equation (4) keeps the semantic representation central and treats the affective resources as controlled corrections. This is often preferable to direct concatenation, which expands the classifier input regardless of resource relevance.

3.2. Feature Construction

Affective features are constructed from four complementary sources and represented as fixed-dimensional vectors at the document level.

For lexical VAD, we compute a three-dimensional vector by averaging valence, arousal, and dominance scores over all tokens covered by the NRC VAD lexicon [2]. Tokens not present in the lexicon are ignored. This produces a transparent but potentially sparse representation, depending on lexical coverage.

For VAD-BERT, the entire input text is processed by a pretrained regression model, and the predicted valence, arousal, and dominance values are used directly as a three-dimensional feature vector [22]. Unlike lexical VAD, this representation is contextual and does not depend on explicit lexicon matches.

For Ekman-style emotion features, we obtain a six-dimensional vector corresponding to basic emotions (anger, disgust, fear, joy, sadness, and surprise). These scores are derived from a pretrained emotion classifier and interpreted as a categorical affective profile.

For SenticNet, we extract concept-level affective information by matching multiword expressions and concepts in the text. The resulting signals are aggregated into a compact vector representation, capturing affective properties beyond individual tokens [7].

All non-textual features are normalized using statistics computed on the training split. This ensures that heterogeneous affective signals are comparable in scale and can be consistently integrated with the transformer representation during fusion.

3.3. Affective Flow

Standard affective representations at the document level collapse the emotional content of a text into a single vector. While this representation is compact and often useful, it ignores the fact that many texts exhibit internal variation in affect as the narrative progresses. Emotional tone may rise, fall, or shift polarity across segments, and these transitions may be informative for downstream classification.

To capture this phenomenon, we define *affective flow* as a trajectory of affective states along the text. Let a document x be divided into K consecutive, non-overlapping segments:

$$\mathcal{S}_K(x) = \{x^{(1)}, x^{(2)}, \dots, x^{(K)}\}, \quad (5)$$

ordered according to their position in the text. For each segment $x^{(k)}$, an affective extractor $g(\cdot)$ produces either a dimensional VAD vector

$$y^{(k)} \in \mathbb{R}^3 \quad (\text{valence, arousal, dominance}) \quad (6)$$

or a contextual embedding

$$e^{(k)} \in \mathbb{R}^{d_v}. \quad (7)$$

The affective flow is then defined as the ordered sequence of affective states together with their first-order differences:

$$F_K = [s^{(1)}, \dots, s^{(K)}, \Delta s^{(2)}, \dots, \Delta s^{(K)}], \quad (8)$$

where $s^{(k)} \in \{y^{(k)}, e^{(k)}\}$ and

$$\Delta s^{(k)} = s^{(k)} - s^{(k-1)}, \quad k = 2, \dots, K. \quad (9)$$

Under this formulation, F_K encodes both the local affective profile of each segment and the transitions between adjacent segments. In the VAD case, this corresponds to modeling how valence, arousal, and dominance evolve across the document. From an interpretive perspective, affective flow can be seen as a path in affective space. A static document-level vector summarizes *where* a text lies in that space, while the flow representation captures *how it moves* through it. Static resources such as SenticNet may achieve higher peak Macro-F1, but EmoFlow introduces an explicit representation of intra-document dynamics; the model is exposed to how affect evolves, not only to where it averages.

Motivation

Figure 1 shows an example of affective flow computed for *Don Quijote de la Mancha*. The example is intended solely to motivate the proposed representation and is independent of the datasets used in the experimental evaluation. This example motivates the design choice behind affective flow representations. A document-level VAD vector collapses the emotional content of a text into a single point, but the curves in Figure 1 reveal clear intra-textual variations that would otherwise be lost.

The text was segmented into consecutive blocks of 100 sentences. For each block, we computed standardized (z-score) VAD values using two independent routes: (i) a lexical VAD approach based on the NRC lexicon, and (ii) an Ekman-category route using a DistilRoBERTa emotion classifier followed by a mapping to VAD space. The resulting curves for valence, arousal, and dominance show how emotional tone evolves across the narrative.

Although the example is literary, this idea applies to other domains. In mental health settings, for instance, therapy transcripts or patient journals may exhibit gradual or abrupt changes in affect. Affective flow provides a compact way to represent these dynamics and may help downstream models distinguish between classes that are not easily separable using static features alone. We formalize a family of affective flow representations derived from VAD-BERT. This formulation generalizes the intuition observed in Figure 1, where segment-level variations reveal patterns that are not captured by a single document-level vector. This extension connects with the general idea of attention and *cross-attention* in

Transformer architectures [11,14], as well as with prior work on gated and multimodal fusion [9,10,17,18]. We now move from this intuition to a formal definition.

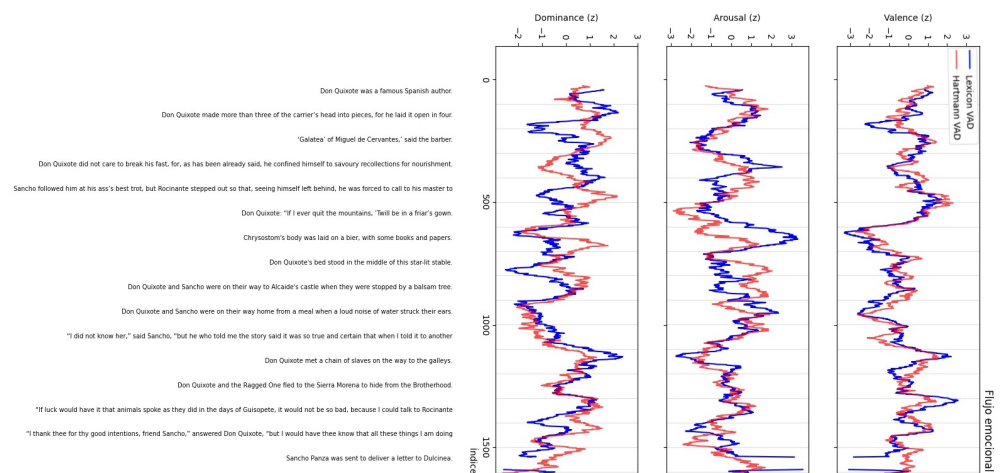


Figure 1. Example of affective flow on *Don Quijote de la Mancha*. Each segment corresponds to a block of 100 sentences. Curves show standardized (z-score) dominance, arousal, and valence. The blue curve is computed using the NRC VAD lexicon, while the red curve is obtained by mapping Ekman-based emotion predictions to VAD space.

Let $x = (w_1, \dots, w_n)$ be the input text and let $h \in \mathbb{R}^d$ be the textual representation obtained from the DistilBERT backbone [15]. To construct the affective flow, we start from a contiguous partition of the text into K segments,

$$\mathcal{S}_K(x) = \{x^{(1)}, x^{(2)}, \dots, x^{(K)}\},$$

where $K \in \{1, 2, 3\}$ depending on the evaluated variant. Whenever possible, the partition respects sentence boundaries; if a robust segmenter is not available, it can be implemented using approximately equiproportional contiguous windows in terms of token count. No strict constraint was imposed on this decision in the present extension.

To handle very short inputs, we keep the dimensionality of EmoFlow fixed for all examples. If a text contains fewer than K words, the available words are repeated minimally until K non-empty segments are obtained. Similarly, if rounding produces an empty segment, the nearest available word is used. This avoids undefined segment-level VAD estimates while preserving a consistent representation across instances.

Each segment $x^{(k)}$ is processed with VAD-BERT, understood here as a contextual regressor that produces a continuous output in the valence–arousal–dominance space [3,22]. For each segment, two types of signals are extracted depending on the variant:

$$y^{(k)} \in \mathbb{R}^3, \quad \text{segment-level VAD output,} \quad (10)$$

$$e^{(k)} \in \mathbb{R}^{d_v}, \quad \text{internal segment embedding.} \quad (11)$$

Prediction-based variants use $y^{(k)}$, while embeddings+deltas variants use $e^{(k)}$ or a combination of both spaces.

To capture affective dynamics, we define first-order differences between adjacent segments:

$$\Delta y^{(k)} = y^{(k)} - y^{(k-1)}, \quad k = 2, \dots, K, \quad (12)$$

$$\Delta e^{(k)} = e^{(k)} - e^{(k-1)}, \quad k = 2, \dots, K. \quad (13)$$

These differences represent local changes in the affective profile along the text. In practice, the evaluated flows are defined as follows:

$$f_{\text{global}} = y_{\text{doc}}, \quad (14)$$

$$f_{\text{flow-2}} = [y^{(1)} \parallel y^{(2)} \parallel \Delta y^{(2)}], \quad (15)$$

$$f_{\text{flow-3}} = [y^{(1)} \parallel y^{(2)} \parallel y^{(3)} \parallel \Delta y^{(2)} \parallel \Delta y^{(3)}]. \quad (16)$$

Under this definition, prediction-space variants have dimensions 3, 9, and 15, respectively. For embedding-based variants, the analogous construction is:

$$f_{\text{flow-3}}^{\text{emb}} = [e^{(1)} \parallel e^{(2)} \parallel e^{(3)} \parallel \Delta e^{(2)} \parallel \Delta e^{(3)}].$$

In the variant referred to as *derived logits*, we use the outputs of the VAD-BERT regression head and their deltas; in many implementations, this output is technically referred to as *logits*, even though it semantically corresponds to continuous VAD predictions.

3.3.1. Pairing Between the Textual Backbone and the Affective Flow

The textual representation h is computed over the full text, while the flow representation is derived from internal segments of the same text. To fuse them, we project the affective signal into the backbone hidden space:

$$z = W_f f + b_f, \quad z \in \mathbb{R}^d.$$

When the fusion module requires a sequence rather than a vector, we use an ordered version of the flow:

$$F_K = [s^{(1)}, \dots, s^{(K)}, \Delta s^{(2)}, \dots, \Delta s^{(K)}] \in \mathbb{R}^{L_K \times d'},$$

where $s^{(k)}$ may be either $y^{(k)}$ or $e^{(k)}$ and $L_K = 1, 3, 5$ for the *global*, *flow-2*, and *flow-3* variants, respectively. Each element of F_K is then projected into the backbone space d before interacting with h .

3.3.2. Evaluated Fusion Mechanisms

Concatenation

The simplest strategy concatenates the textual representation with the flow representation:

$$h_{\text{concat}} = [h \parallel f].$$

This variant serves as a reference to assess whether affective dynamics provide useful information even without explicit relevance control.

Scalar Gating

The *scalar gating* variant follows the principle of the main paper, now applied to the VAD-BERT flow:

$$z = W_f f + b_f, \quad (17)$$

$$\alpha = \sigma(w_\alpha^\top [h \parallel z] + c_\alpha), \quad (18)$$

$$h_{\text{sg}} = h + \alpha z, \quad (19)$$

where $\alpha \in (0, 1)$ is a scalar per instance. This formulation learns how much the affective flow should influence the final representation.

Dedicated VAD Gating (BertWithVADGating)

In this variant, the scalar is replaced with a vector gate:

$$z = W_f f + b_f, \quad (20)$$

$$g = \sigma(W_g[h \parallel z] + b_g), \quad g \in \mathbb{R}^d, \quad (21)$$

$$h_{vg} = h + g \odot z, \quad (22)$$

where $\odot$ denotes element-wise multiplication. This variant offers greater capacity, although it may overfit when using low-dimensional auxiliary signals.

Dedicated Scalar VAD Gate (BertWithVADScalarGate)

This variant preserves the dedicated VAD architecture but uses a scalar gate:

$$z = W_f f + b_f, \quad (23)$$

$$\alpha = \sigma(w_\alpha^\top [h \parallel z] + c_\alpha), \quad (24)$$

$$h_{vsg} = h + \alpha z. \quad (25)$$

In practice, this formulation helps isolate whether improvements depend on the dedicated architecture or on the use of a scalar gate.

Cross-Attention Fusion

To evaluate more structured interactions between text and flow, we introduce a cross-attention variant. The textual representation serves as the query, whereas the affective flow states are projected into keys and values through trainable linear transformations:

$$q = W_q h, \quad (26)$$

$$K = F_K W_k, \quad (27)$$

$$V = F_K W_v, \quad (28)$$

$$a = \text{softmax}\left(\frac{qK^\top}{\sqrt{d_k}}\right), \quad (29)$$

$$c = aV, \quad (30)$$

$$h_{ca} = \text{LN}(h + c). \quad (31)$$

where W_q , W_k , and W_v are trainable projection matrices learned jointly with the classifier during end-to-end optimization. $W_q \in \mathbb{R}^{d \times d_k}$, $W_k \in \mathbb{R}^{d' \times d_k}$, and $W_v \in \mathbb{R}^{d' \times d}$ are trainable projection matrices, with d_k denoting the attention key dimension. The intuition is that the model can decide not only *how much* affective signal to use, but also *which part of the flow* is most informative.

Cross-Attention with Residual Scalar Control (cross_attention_alpha)

In the `cross_attention_alpha` variant, the attended context c is reinjected using a learned scalar:

$$\beta = \sigma(w_\beta^\top [h \parallel c] + c_\beta), \quad (32)$$

$$h_{ca-\alpha} = h + \beta c. \quad (33)$$

This formulation aims to combine the relational benefit of attention with the regularizing bias of a scalar gate.

3.3.3. Classification, Loss, and Normalization

The fused representation h' is passed to a linear classification layer:

$$\hat{y} = \text{Classifier}(h').$$

All experiments reported in this work use a weighted cross-entropy loss to compensate for class imbalance. The class weights are computed from the training partition and remain fixed throughout training.

All non-textual features are normalized using the mean and standard deviation computed on the training set. For the VAD-BERT variants, the auxiliary VAD regressor is kept frozen and used exclusively as a fixed feature extractor, allowing the evaluation to focus on the proposed fusion mechanisms rather than on additional adaptation of the affective encoder.

Unless otherwise stated, all experiments use five random seeds, three training epochs, a learning rate of 2×10^{-5} , class-weighted optimization, and paired bootstrap significance testing.

3.3.4. Effect of Scalar Gating on the Decision Function

The scalar gate can be interpreted as an instance-dependent modulation of the classifier input. Let h denote the textual representation produced by DistilBERT and let z denote the projected affective feature vector. The fused representation is

$$h' = h + \alpha z,$$

where $\alpha \in (0, 1)$ is predicted from the input itself.

Unlike simple concatenation, where the auxiliary features are always presented to the classifier, scalar gating makes the contribution of the affective signal conditional on each instance. Geometrically, the gate moves the representation from h toward the affective direction z , with the displacement controlled by α . When α is close to zero, the classifier behaves similarly to the text-only model. As α increases, the decision function becomes progressively more influenced by the affective representation.

Consequently, the proposed mechanism does not introduce a new classifier or a different decision rule. Instead, it adaptively modifies the representation on which the same classifier operates. This interpretation also explains why the learned gate values provide a direct indication of when the model relies on affective information, making the fusion process inherently interpretable.

4. Experimental Setup

This section describes the experimental configuration used to evaluate the proposed fusion methods and affective flow representations. The goal is to ensure a controlled comparison across variants while maintaining reproducibility.

4.1. Model and Training Configuration

All experiments use `distilbert-base-uncased` as the textual backbone. Unless otherwise stated, models are trained for three epochs using the Adam optimizer with a learning rate of 2×10^{-5} . Class weights are applied to mitigate label imbalance, and performance is evaluated over five random seeds. Statistical significance is assessed using paired bootstrap with 3000 samples.

All experiments were conducted on an NVIDIA L4 GPU in Google Colab using automatic mixed-precision training. Given the relatively compact DistilBERT architecture,

the maximum sequence lengths considered in this work, and the use of mixed precision, the reported batch sizes could be used directly without gradient accumulation.

4.2. Datasets

We evaluate the proposed framework on two complementary datasets that represent distinct affective text classification scenarios. The MentalHealth dataset focuses on domain-specific mental health detection, whereas GoEmotions is a large-scale benchmark for fine-grained emotion recognition in open-domain text. Together, they allow us to assess the proposed fusion strategies across substantially different linguistic and semantic conditions.

These datasets also exhibit complementary strengths and limitations. The MentalHealth dataset represents a clinically relevant application but is relatively specialized, whereas GoEmotions provides broader linguistic diversity but is derived from Reddit discussions and therefore reflects a particular communication style. Although consistent improvements across both datasets suggest that the proposed fusion strategy is not tied to a single corpus, further evaluation on additional domains, languages, and text classification tasks would be needed to establish its broader generalizability.

MentalHealth Corpus Setup

For the MentalHealth experiments, we use a maximum sequence length of 128 and a batch size of 512. The label set consists of *Anxiety*, *Bipolar*, *Depression*, *Normal*, *Personality disorder*, *Stress*, and *Suicidal*.

The comparison includes text-only baselines, concatenation-based fusion, scalar-gated fusion, and a vectorial gating variant. All affective features are normalized using training-set statistics prior to fusion.

GoEmotions Setup

For GoEmotions, we use a maximum sequence length of 32. The DistilBERT batch size is 512, while the VAD-BERT batch size is set to 128 due to its separate inference pipeline. All other hyperparameters match the MentalHealth configuration.

In this corpus, we compare text-only models with concatenation and scalar gating for lexical VAD, VAD-BERT, SenticNet, and selected combinations.

4.3. Evaluation Metrics

Macro-F1 is the central metric because both corpora are label-imbalanced and the research question focuses on balanced class performance rather than dominant-class accuracy alone. We also report accuracy for completeness, and for GoEmotions we include weighted F1 to analyze whether improvements reflect better minority-class behavior or simply majority-class performance.

Unless otherwise noted, tables report mean performance over five random seeds. Statistical variability is summarized either as standard deviation or, when paired comparisons are the focus, by bootstrap confidence intervals and significance tests.

5. Results and Discussion

This section presents the empirical results in three stages. First, we analyze the MentalHealth corpus, where affective information acts as an auxiliary signal for mental health-related text classification. Second, we examine GoEmotions, where affect is the primary target and the label space is substantially more fine-grained. Third, we study the extended set of affective-flow variants derived from VAD-BERT, including EmoFlow representations, scalar gating, dedicated gating, and cross-attention. Across the section, Macro-F1 is treated as the main metric because both datasets are label-imbalanced and because the purpose of

affective fusion is not only to improve dominant-class behavior, but also to support more balanced classification.

5.1. MentalHealth Results

Table 1 summarizes the main MentalHealth results. The strongest result is obtained by *ScalarGate* (VAD + Ekman + SenticNet), which reaches 0.7480 Macro-F1. The best concatenation model, using the same three resources, reaches 0.7472 Macro-F1 and obtains the highest accuracy, 0.7791. Although the difference in Macro-F1 between these systems is small, the overall comparison across matched feature sets provides a clearer picture.

Table 1. MentalHealth results. Higher is better. Best values for Macro-F1 and Accuracy are shown in bold.

Variant	Fusion	Macro-F1	Accuracy	Δ vs. Text
ScalarGate (VAD + Ekman + SenticNet)	gating	0.7480	0.7786	+0.0072
Text + VAD + Ekman + SenticNet	concatenation	0.7472	0.7791	+0.0064
ScalarGate (VAD)	gating	0.7468	0.7782	+0.0060
Text + Ekman + SenticNet	concatenation	0.7446	0.7774	+0.0038
ScalarGate (Ekman + SenticNet)	gating	0.7454	0.7771	+0.0046
ScalarGate (Ekman)	gating	0.7454	0.7784	+0.0046
Text + VAD	concatenation	0.7429	0.7761	+0.0021
Text + Ekman	concatenation	0.7422	0.7740	+0.0014
ScalarGate (VAD + Ekman)	gating	0.7444	0.7766	+0.0036
ScalarGate (SenticNet)	gating	0.7433	0.7771	+0.0025
Text + VAD + SenticNet	concatenation	0.7420	0.7744	+0.0012
ScalarGate (VAD + SenticNet)	gating	0.7423	0.7761	+0.0015
BertWithVADGating (vectorial)	gating	0.7414	0.7747	+0.0006
Text + SenticNet	concatenation	0.7412	0.7760	+0.0004
Text only (baseline)	—	0.7408	0.7730	0.0000

Across directly comparable configurations, scalar gating tends to outperform concatenation, both in single-resource settings (e.g., VAD, Ekman) and in multi-resource combinations (e.g., VAD + SenticNet, Ekman + SenticNet). This pattern suggests that the gain is not tied to a particular feature configuration, but to the way affective information is integrated into the representation.

The VAD-only comparison is particularly informative. *ScalarGate* (VAD) reaches 0.7468 Macro-F1, whereas *Text* + VAD reaches 0.7429. The same affective signal is therefore more effective when its influence is modulated rather than directly appended. A similar tendency appears for Ekman-style features and SenticNet, although the magnitude of the improvement varies across resources.

The multi-resource results indicate that adding more affective information is not sufficient on its own. While the combination of VAD, Ekman, and SenticNet yields the best Macro-F1 under scalar gating, simpler configurations remain competitive. For instance, *ScalarGate* (VAD) performs close to the top model despite using a single resource. This suggests that the benefit of richer combinations depends on controlled integration rather than feature accumulation alone.

The vectorial gating baseline, *BertWithVADGating*, provides a useful contrast. Its Macro-F1 (0.7414) is only slightly above the text-only baseline and clearly below the scalar-gated variants. This supports the use of a single scalar coefficient for compact affective features. A dimension-wise gate introduces additional capacity, but does not appear beneficial in this setting and may increase sensitivity to noise.

Accuracy and Macro-F1 emphasize different aspects of performance. The best concatenation model achieves the highest accuracy, while the best scalar-gated model achieves the highest Macro-F1. This difference reflects the role of class imbalance: accuracy is driven by dominant classes, whereas Macro-F1 gives equal weight to all labels. Since affective

features are expected to support more difficult or minority categories, Macro-F1 is the more informative metric in this context.

Figure 2 complements the aggregate results by showing the distribution of Macro-F1 across seeds. Beyond the top-line values, the relevant question is whether a method remains competitive across runs. In this respect, the scalar-gated variants show stable behavior, reinforcing the interpretation that their advantage is not driven by isolated favorable seeds.

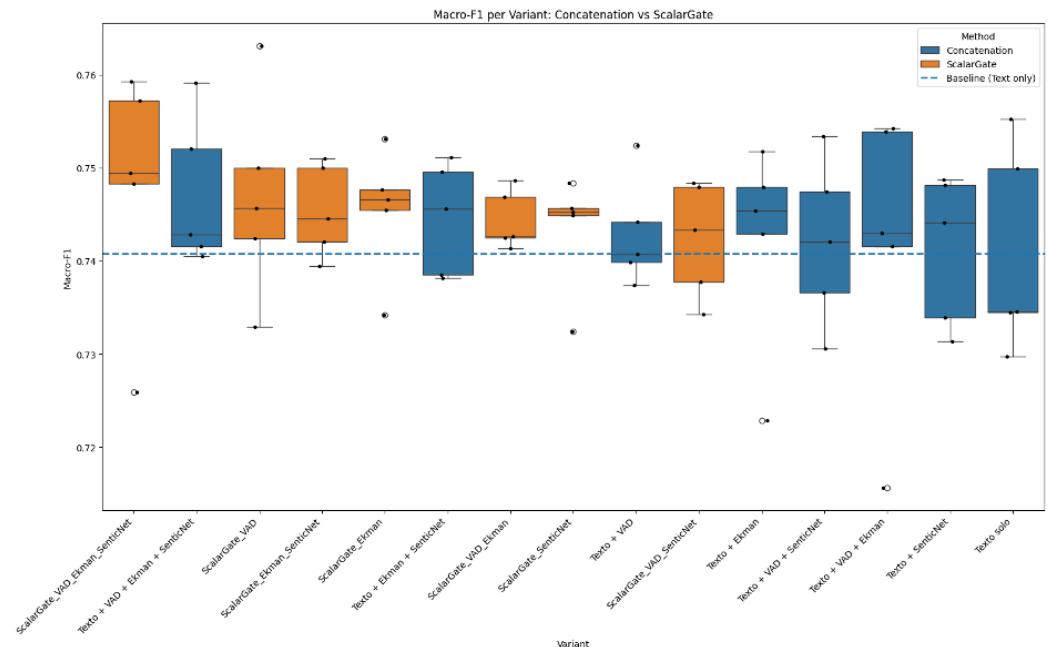


Figure 2. Boxplot of seed-wise Macro-F1 on the MentalHealth corpus. Boxes indicate the interquartile range (IQR), the horizontal line inside each box represents the median, and whiskers extend to 1.5 times the IQR. Black dots show the individual seed-wise results, with a small horizontal jitter added solely for visualization. Open circles indicate outlier observations according to the boxplot criterion.

5.2. GoEmotions Results

Table 2 reports the GoEmotions multi-seed summary. The GoEmotions results are more compressed than the MentalHealth results. Most variants lie within a narrow Macro-F1 interval, from 0.3891 for the text-only baseline to 0.3932 for the best model. This small spread indicates that the task is difficult and that the DistilBERT baseline already captures much of the information needed for prediction. Still, the ranking is informative.

Table 2. GoEmotions multi-seed summary. Mean $\pm$ standard deviation over five seeds.

Variant	Fusion	Test Macro-F1	Test Accuracy	Weighted F1
Text + VAD-BERT	scalar_gate	0.3932 $\pm$ 0.0032	0.4122 $\pm$ 0.0043	0.3956 $\pm$ 0.0035
Text + SenticNet	scalar_gate	0.3929 $\pm$ 0.0053	0.4157 $\pm$ 0.0067	0.4023 $\pm$ 0.0082
Text + VAD-BERT + SenticNet	scalar_gate	0.3918 $\pm$ 0.0083	0.4102 $\pm$ 0.0092	0.3926 $\pm$ 0.0099
Text + VAD lexical	scalar_gate	0.3911 $\pm$ 0.0059	0.4113 $\pm$ 0.0085	0.3974 $\pm$ 0.0102
Text + VAD-BERT + SenticNet	concatenation	0.3910 $\pm$ 0.0028	0.4115 $\pm$ 0.0033	0.3963 $\pm$ 0.0071
Text + VAD lexical + SenticNet	concatenation	0.3910 $\pm$ 0.0063	0.4119 $\pm$ 0.0054	0.3963 $\pm$ 0.0049
Text + VAD lexical + SenticNet	scalar_gate	0.3902 $\pm$ 0.0026	0.4121 $\pm$ 0.0060	0.3955 $\pm$ 0.0086
Text + VAD lexical	concatenation	0.3902 $\pm$ 0.0044	0.4112 $\pm$ 0.0051	0.3951 $\pm$ 0.0067
Text + VAD-BERT	concatenation	0.3901 $\pm$ 0.0049	0.4121 $\pm$ 0.0076	0.3967 $\pm$ 0.0085
Text + SenticNet	concatenation	0.3900 $\pm$ 0.0053	0.4099 $\pm$ 0.0050	0.3922 $\pm$ 0.0034
Text only	baseline	0.3891 $\pm$ 0.0060	0.4104 $\pm$ 0.0065	0.3941 $\pm$ 0.0076

The best Macro-F1 is obtained by scalar-gated *Text + VAD-BERT*, with 0.3932. The corresponding concatenation model reaches 0.3901. This difference is small in absolute terms, but it is consistent with the central hypothesis of the paper: contextual affective information is more useful when its contribution is controlled than when it is appended unconditionally.

The scalar-gated *Text + SenticNet* model obtains the best accuracy and weighted F1. This suggests that SenticNet may be especially useful for the most frequent or easier-to-classify labels, since weighted F1 gives more influence to classes with more examples. In contrast, VAD-BERT obtains the best Macro-F1, which indicates a slightly better balance across the full label set. This difference matters because GoEmotions contains many fine-grained and imbalanced categories. A model that improves weighted F1 is not necessarily the same model that improves balanced class performance.

The multi-resource combinations do not dominate the single-resource variants. *Text + VAD-BERT + SenticNet* with scalar gating reaches 0.3918 Macro-F1, below both scalar-gated *Text + VAD-BERT* and scalar-gated *Text + SenticNet*. This result suggests that combining affective resources can introduce redundancy or noise, especially when the label space is fine-grained and the auxiliary signals are not aligned in the same way. In this setting, a cleaner single-resource signal may be easier for the gate to exploit.

Scalar gating improves four of the five directly matched comparisons: VAD-BERT, SenticNet, VAD-BERT + SenticNet, and lexical VAD. The only exception is the lexical VAD + SenticNet combination, where concatenation slightly outperforms scalar gating. This exception is useful because it prevents an overly broad conclusion. Scalar gating is generally better in these experiments, but it is not universally superior. Its benefit depends on the interaction between the feature family, the label structure, and the amount of useful information already captured by the text encoder.

The standard deviations also deserve attention. Some scalar-gated variants, such as *Text + VAD lexical + SenticNet*, show relatively low standard deviation in Macro-F1, whereas others, such as *Text + VAD-BERT + SenticNet*, are more variable. This means that the highest mean is not the only criterion for evaluating a variant. Stability across seeds is particularly important when the observed gains are small, as they are in GoEmotions.

Another detail is the selected best epoch. The scalar-gated top variants tend to peak around the final epoch, while several concatenation runs peak slightly earlier. This is consistent with the interpretation that gating introduces a more selective fusion process. The model may require slightly more training to learn when to use the affective signal, but this same selectivity may prevent the auxiliary features from being amplified too early.

5.3. Comparison Across MentalHealth and GoEmotions

The two corpora show related but not identical patterns. On MentalHealth, scalar gating produces a clearer and more uniform advantage over concatenation. On GoEmotions, the advantage is smaller and more dependent on the resource. This difference is expected. In MentalHealth classification, affective information is an auxiliary cue that may help distinguish overlapping categories such as anxiety, stress, depression, and suicidal discourse. In GoEmotions, affect is already the target label space, so the textual encoder may already encode much of the relevant emotional information.

The results also suggest that the usefulness of a resource depends on the evaluation metric. VAD-BERT is especially competitive in Macro-F1, whereas SenticNet often performs strongly in accuracy and weighted F1. This pattern supports the idea that contextual dimensional affect and concept-level affective commonsense contribute different kinds of information. VAD-BERT may help distribute performance more evenly across labels, while SenticNet may better support frequent or semantically salient classes.

Across both datasets, concatenation remains a strong baseline. It is simple, stable, and often close to the best model. However, it has no mechanism for deciding whether an affective feature is useful for a particular input. Scalar gating provides this capability while introducing only a small number of additional trainable parameters, consisting of a projection layer and a scalar gating function for each affective resource. Although the empirical gains are modest, they are remarkably consistent across datasets, affective resources, and matched comparisons. More importantly, the scalar gate offers an interpretable instance-level mechanism that explicitly reveals when external affective information contributes to the prediction. From this perspective, the practical value of the proposed fusion lies not only in its predictive performance, but also in its lightweight design and its ability to analyze the role of affective information during inference.

This conclusion should remain measured. Scalar gating is not a radical architectural change, and it does not transform the task. Its value lies in providing a small instance-wise control mechanism that helps the model avoid treating all affective side information as equally reliable. In heterogeneous affective settings, that small amount of control appears useful.

5.4. Affective Flow Representations

To assess whether differences between variants were statistically meaningful, we performed a paired bootstrap hypothesis test by comparing each model against the text-only baseline over the test set. The test statistic was the paired difference in Macro-F1 between the evaluated variant and the baseline, computed for each bootstrap replicate. The null hypothesis (H_0) states that the expected Macro-F1 difference between a given variant and the baseline is zero, whereas the alternative hypothesis (H_1) states that the difference is non-zero. For each bootstrap replicate, the paired test instances were resampled with replacement, the difference in Macro-F1 was recomputed, and the empirical distribution of these differences was used to estimate both the 95% confidence interval and the corresponding p-value. Statistical significance was determined at the $\alpha = 0.05$ level: improvements whose confidence intervals excluded zero and whose p-values were below 0.05 were considered statistically significant.

Table 3 extends the comparison with VAD-BERT-based affective flow variants. For descriptive purposes, all variants are compared against a common *Text only* baseline with Macro-F1 = 0.628322. The table includes static affective resources, VAD-BERT global features, EmoFlow-2 and EmoFlow-3 variants, dedicated gating mechanisms, cross-attention mechanisms, and two negative variants that use either derived logits or embeddings with deltas. While EmoFlow does not dominate the absolute ranking, it provides a structurally richer and statistically more reliable signal within the VAD-BERT family. This shifts the evaluation from isolated point estimates to consistency across resamples and configurations.

In Table 3, Δ denotes the difference in Macro-F1 with respect to the text-only baseline. Positive values indicate an improvement, whereas negative values indicate a degradation. The 95% confidence interval summarizes the uncertainty of this difference across bootstrap samples. When the interval excludes zero, the improvement (or degradation) is considered statistically significant. The reported p-value corresponds to the probability of observing a difference at least as extreme under the null hypothesis. Consequently, variants marked as “Yes” in the significance column provide statistically significant evidence of improvement over the baseline, whereas variants marked as “No” do not provide sufficient evidence despite occasionally exhibiting higher point estimates. The highest absolute Macro-F1 in Table 3 is obtained by *Text + SenticNet*, with 0.6501. It is followed by *Text + VAD + SenticNet* and *Text + Ekman*. This means that the best static lexical or concept-level resources remain very competitive. EmoFlow representations do not replace these resources in raw

performance. Instead, their contribution becomes clearer when the comparison is restricted to the VAD/VAD-BERT family.

Table 3. Comparison of affective variants (sorted by Macro-F1). Abbreviations: T = Text, V = VAD, S = SenticNet, E = Ekman, VB = VAD-BERT, F2/F3 = EmoFlow-2/EmoFlow-3, G = global, SG = scalar gating, DG = dedicated gating, D-SG = dedicated scalar gating, CA = cross-attention, CA- α = cross-attention with scalar control, DL = derived logits, ED = embeddings + deltas.

Rank	Variant	Fusion	Macro-F1	Acc.	W-F1	Δ	95% CI	p	Sig.
1	T + S	Concat	0.6501	0.7053	0.7070	+0.0217	[0.0056, 0.0234]	0.0020	Yes
2	T + V + S	Concat	0.6488	0.7035	0.7088	+0.0205	[0.0053, 0.0207]	0.0007	Yes
3	T + E	Concat	0.6477	0.7034	0.7093	+0.0193	[0.0046, 0.0196]	0.0007	Yes
4	T + VB(F3)-DG	DG	0.6460	0.6994	0.7075	+0.0177	[−0.0029, 0.0136]	0.1747	No
5	T + VB(F3)-CA	CA	0.6457	0.7010	0.7024	+0.0174	[0.0081, 0.0269]	<1/ N_{boot}	Yes
6	T + VB(F3)-D-SG	D-SG	0.6453	0.7028	0.7100	+0.0170	[−0.0031, 0.0131]	0.2473	No
7	T + VB(F3)-SG	SG	0.6453	0.6987	0.7050	+0.0169	[0.0078, 0.0255]	<1/ N_{boot}	Yes
8	T + V + E	Concat	0.6446	0.6971	0.7048	+0.0163	[0.0016, 0.0165]	0.0233	Yes
9	T + V	Concat	0.6431	0.6991	0.7063	+0.0148	[−0.0001, 0.0154]	0.0553	No
10	T + E + S	Concat	0.6411	0.6990	0.7080	+0.0128	[−0.0029, 0.0135]	0.2067	No
11	T + V + E + S	Concat	0.6405	0.6986	0.7071	+0.0122	[−0.0027, 0.0127]	0.2320	No
12	T + VB(G)-SG	SG	0.6403	0.6967	0.7010	+0.0120	[0.0028, 0.0213]	0.0107	Yes
13	T + VB(F2)-SG	SG	0.6384	0.6952	0.7053	+0.0101	[0.0016, 0.0186]	0.0167	Yes
14	T (baseline)	—	0.6283	0.6847	0.6909	0.0000	—	1.0000	No
15	T + VB(F3)-DL	Concat	0.5915	0.6658	0.6720	−0.0368	[−0.0482, −0.0256]	<1/ N_{boot}	Worse
16	T + VB(F3)-ED-CA α	CA- α	0.5896	0.6673	0.6799	−0.0387	[−0.0496, −0.0279]	<1/ N_{boot}	Worse

Within the VAD-BERT family, EmoFlow-3 is the strongest representation. The cross-attention variant reaches 0.6457 Macro-F1, while the scalar-gated variant reaches 0.6453. Both are above the VAD-BERT global scalar-gated representation, which reaches 0.6403, and above EmoFlow-2, which reaches 0.6384. This ranking suggests that modeling multiple affective states and their transitions can provide useful information beyond a single document-level VAD prediction.

The difference between EmoFlow-2 and EmoFlow-3 is also meaningful. EmoFlow-2 already improves over the baseline and has a positive confidence interval. However, EmoFlow-3 achieves a stronger Macro-F1 and a larger estimated improvement. This suggests that a minimal two-state trajectory may capture some affective change, but a three-state trajectory provides a richer description of intra-document affective progression.

The strongest EmoFlow models do not achieve the highest absolute Macro-F1 in the table, but they are statistically reliable. Both $T+VB(F3)-CA$ and $T+VB(F3)-SG$ have confidence intervals fully above zero. This indicates that their improvements over the baseline are not driven only by a favorable point estimate. The forest plot in Figure 3 makes this especially visible: the confidence intervals for the strongest EmoFlow-3 variants remain entirely on the positive side of the baseline line.

A key observation is that the strongest EmoFlow-3 variants combine competitive mean performance with statistically reliable improvements. In contrast, several alternatives with similar or even slightly higher point estimates exhibit confidence intervals that cross zero. This indicates that stability, rather than isolated gains, is the defining advantage of EmoFlow-based representations.

The comparison between $T+VB(F3)-CA$, $T+VB(F3)-SG$, $T+VB(F3)-DG$, and $T+VB(F3)-D-SG$ is especially instructive. All four have similar mean Macro-F1 values, but only the cross-attention and standard scalar-gated variants are significant. The dedicated VAD gating and dedicated scalar-gating variants have intervals that cross zero. This shows that statistical reliability is not determined only by the mean improvement. Stability across bootstrap resamples matters as much as the point estimate.

The dedicated vectorial gate is particularly revealing. Although it reaches 0.6460 Macro-F1, its confidence interval crosses zero. This suggests that its higher point estimate should be interpreted cautiously. The additional capacity of a vector gate may make the

model more flexible, but it may also make the effect less stable. In contrast, the scalar gate is less expressive but more regularized, which is consistent with the central claim of the paper.

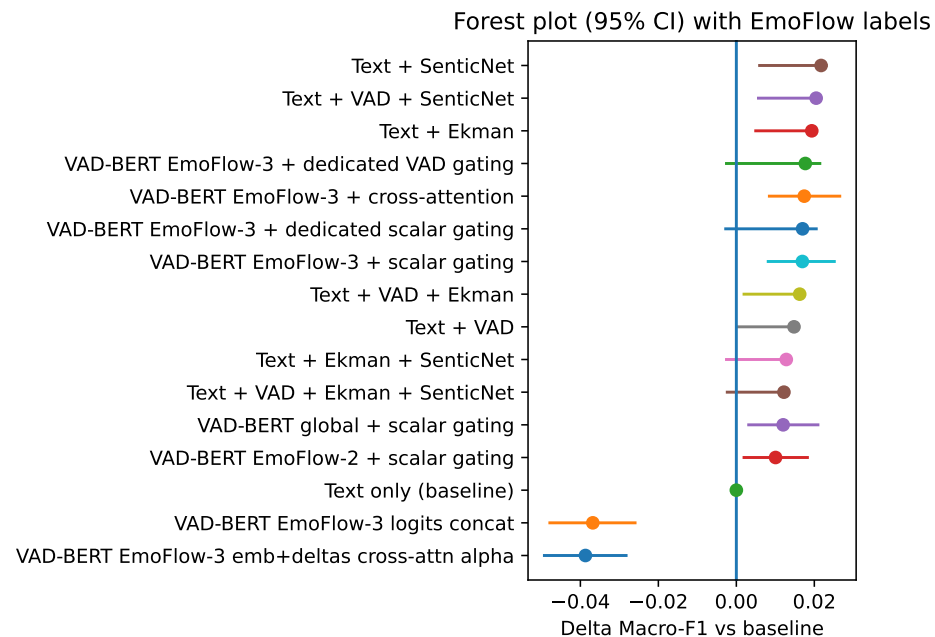


Figure 3. Forest plot of Δ Macro-F1 vs. baseline with 95% confidence intervals for static affective resources, VAD-BERT global features, EmoFlow variants, and negative fusion variants.

The static lexical variants also deserve attention. *Text + VAD* reaches 0.6431 Macro-F1 but has a confidence interval that narrowly crosses zero and a p -value of 0.0553. This is an interesting borderline result. It suggests that lexical VAD may help, but the evidence is less stable than for the best EmoFlow-3 variants. By contrast, *Text + SenticNet*, *Text + VAD + SenticNet*, and *Text + Ekman* all show positive confidence intervals and statistically significant improvements.

There are also informative negative results. The derived-logits variant with concatenation reaches only 0.5915 Macro-F1, and the embeddings+deltas variant with cross-attention alpha reaches 0.5896. Both are significantly worse than the baseline. These results are useful because they show that dynamic affective information is not automatically beneficial. Poorly controlled or poorly aligned fusion can damage the textual representation. In other words, the benefit comes not from adding more affective signals, but from integrating them in a way that preserves the backbone representation while selectively injecting affective structure.

The forest plot reinforces this interpretation. Positive and reliable variants appear clearly to the right of zero, unreliable variants overlap the baseline line, and the negative variants lie entirely to the left. This visualization therefore summarizes the central empirical message of the extended experiment: affective information helps only when its integration is stable and controlled. EmoFlow-3 with scalar gating and cross-attention satisfies this condition; the derived-logits and embeddings+deltas variants do not.

5.5. Methodological Interpretation

The results point to a distinction between performance and structure. The highest absolute scores are still achieved by strong static affective resources, particularly SenticNet and its combinations. However, the EmoFlow experiments show that dynamic representations provide a consistent and interpretable signal within the VAD-BERT family. This shifts the contribution of the paper from feature comparison to integration strategy.

Two complementary questions emerge. Scalar gating addresses *when* affective information should influence the model, introducing an instance-level control over auxiliary signals. EmoFlow addresses *how* affect evolves within the text, capturing transitions that cannot be represented by a single document-level vector. Static representations summarize affective content, whereas flow-based representations model its internal dynamics. Both require controlled integration to avoid introducing noise into the textual backbone.

Under this perspective, the title claim, *Learning When to Feel*, remains appropriate. The results suggest that the model benefits from learning not only when to use affective information, but also how to exploit its internal evolution. This distinction is modest in terms of performance gains, but methodologically relevant.

5.6. Limitations of the Present Results

Several limitations should be considered. For example, current results measure robustness under resampling of the same test set, not robustness under domain shift. A positive bootstrap interval indicates stability with respect to the evaluated sample, but it does not guarantee that the same variant will generalize to a new corpus, platform, or annotation scheme.

Second, the EmoFlow variants depend on segmentation. The present formulation uses a small number of segments, but other choices are possible: sentence-based segmentation, token-window segmentation, overlapping windows, or discourse-informed segmentation. The quality of the affective trajectory may depend on this design decision.

Third, the negative results indicate that not all dynamic representations are useful. Embeddings, deltas, logits, and cross-attention variants can introduce noise if they are not properly aligned with the textual representation. Future work should therefore include a more systematic ablation of affective levels versus deltas, prediction-space flows versus embedding-space flows, and scalar versus vector control.

Finally, the observed gains are small in absolute terms. This does not make them irrelevant, but it does require caution when interpreting their practical impact. Rather than proposing a substantially more complex architecture to maximize performance, this work explores a lightweight and interpretable fusion mechanism whose additional computational cost is minimal. We therefore view the principal contribution of the proposed approach as methodological: understanding when and how affective information should be incorporated into transformer-based classifiers.

6. Conclusions

This article studied how external affective information should be integrated into a compact transformer-based text classifier. Rather than treating affective resources as features to be appended mechanically, we examined whether their contribution should be controlled through learned fusion mechanisms. The comparison was carried out on two corpora with different target structures: MentalHealth, where affective cues support the classification of mental health-related statements, and GoEmotions, where affect is itself the target of prediction.

The first main finding is that scalar-gated fusion is a strong and reliable alternative to plain concatenation. On MentalHealth, scalar gating obtains the best Macro-F1 and outperforms concatenation in all directly matched feature configurations. On GoEmotions, the gains are smaller, but scalar gating still gives the best Macro-F1 and improves most matched comparisons. These results suggest that even a very simple gate can be useful when affective resources are compact, heterogeneous, and only partially reliable.

The second finding is that more affective information is not automatically better. Some combinations of resources improve performance, but others provide only marginal gains

or introduce instability. This is especially clear in the comparison between static lexical or concept-level resources and VAD-BERT-derived representations. SenticNet and some lexical combinations remain highly competitive in absolute Macro-F1, which indicates that carefully curated affective commonsense still provides a strong signal. At the same time, VAD-BERT-based representations offer a different advantage: they are contextual and can be used to model affect beyond direct lexical matches.

The extended EmoFlow experiments refine this picture. Affective flow representations do not simply replace static affective features, nor do they dominate the full table in raw performance. Their value lies elsewhere. Within the VAD-BERT family, EmoFlow-3 variants provide stronger and more stable improvements than global VAD-BERT representations and EmoFlow-2. In particular, EmoFlow-3 with scalar gating and EmoFlow-3 with cross-attention show statistically reliable gains over the baseline. This supports the idea that affective dynamics can contribute useful structure when they are integrated through controlled fusion mechanisms.

The negative results are also important. Some variants based on derived logits or embeddings with deltas perform substantially below the baseline. This shows that dynamic affective information is not beneficial by default. If the auxiliary signal is poorly aligned with the textual representation, or if the fusion mechanism lacks sufficient control, affective features can degrade rather than improve the classifier. This reinforces the central argument of the paper: the question is not only whether affective information is available, but how it is injected into the model.

Taken together, the results support a restrained but meaningful conclusion. Scalar gating is not a radical architectural intervention, and the gains are not large enough to claim that it transforms the task. Its value is methodological: it gives the model a lightweight instance-wise mechanism for deciding when affective information should matter. EmoFlow extends this idea by asking how affect changes within the text. In this sense, the title claim, *Learning When to Feel*, can be understood in two complementary ways: learning when to use affective side information, and learning how to exploit its internal evolution.

Several limitations remain. Some comparisons in the extended table are descriptive and should be interpreted cautiously unless all confidence intervals and paired bootstrap tests are recomputed against the same baseline predictions. The present results also evaluate stability under resampling of the same test set, not robustness under domain shift. In addition, EmoFlow depends on segmentation choices, and the current experiments only explore a small set of variants. More systematic ablations are needed to isolate the role of affective levels, deltas, segment length, contextual VAD prediction, and attention-based fusion.

Future work should therefore move in four directions. First, paired tests should be run directly between EmoFlow-3, lexical VAD, SenticNet, and their hybrid combinations. Second, flow representations should be evaluated under cross-domain and cross-corpus settings, especially from GoEmotions to mental health-related classification. Third, the gate values and attention weights should be analyzed to determine when the model actually relies on affective information. Fourth, hybrid static–dynamic representations should be explored, combining concept-level affective commonsense with contextual affective trajectories.

Overall, the evidence suggests that affective resources remain useful even in the presence of transformer encoders, but their usefulness depends strongly on the fusion strategy. Plain concatenation is a strong baseline, yet it treats all auxiliary information as equally relevant. Scalar gating and EmoFlow offer a more disciplined alternative: they keep the textual backbone central while allowing affective information to enter selectively and, when appropriate, dynamically. This provides a practical and interpretable direction for affect-enriched NLP models.

Author Contributions: Conceptualization, H.C. and M.H.L.; methodology, H.C. and M.H.L.; software, H.C. and M.H.L.; validation, H.C., P.G. and G.M.; formal analysis, H.C. and G.M.; investigation, H.C., M.H.L. and P.G.; resources, M.H.L., P.G. and G.M.; writing—original draft preparation, H.C. and M.H.L.; writing—review and editing, H.C., M.H.L. and P.G.; supervision, H.C., P.G. and G.M.; project administration, H.C., P.G. and G.M.; funding acquisition, H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Mexican Government through Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI), Comisión de Operación y Fomento de Actividades Académicas del IPN (COFAA-IPN), and Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico, under Grant 2026-0569.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors wish to thank SECIHTI and COFAA-IPN (Instituto Politécnico Nacional) for their invaluable support for developing this research.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BCE	Binary Cross-Entropy
BERT	Bidirectional Encoder Representations from Transformers
CA	Cross-Attention
CA- α	Cross-Attention with scalar control
CI	Confidence Interval
DG	Dedicated (vector) gating
D-SG	Dedicated scalar gating
DL	Derived logits
ED	Embeddings and deltas
EmoFlow	Affective flow representation
F1	F1-score
F2	EmoFlow with 2 segments
F3	EmoFlow with 3 segments
G	Global (document-level representation)
LN	Layer Normalization
LR	Learning Rate
ML	Machine Learning
NLP	Natural Language Processing
PAD	Pleasure–Arousal–Dominance (equivalent to VAD)
SG	Scalar Gating
T	Text (textual backbone only)
V	VAD (Valence, Arousal, Dominance)
VB	VAD-BERT
W-F1	Weighted F1-score
Δ	Difference with respect to baseline

References

1. Ekman, P. An Argument for Basic Emotions. *Cogn. Emot.* **1992**, *6*, 169–200. [[CrossRef](#)]
2. Mohammad, S.M. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Kerrville, TX, USA, 2018; pp. 174–184. [[CrossRef](#)]

3. Buechel, S.; Hahn, U. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2, Short Papers), Valencia, Spain, 3–7 April 2017*; Association for Computational Linguistics: Kerrville, TX, USA, 2017; pp. 578–585.
4. Landowska, A. Towards New Mappings between Emotion Representation Models. *Appl. Sci.* **2018**, *8*, 274. [\[CrossRef\]](#)
5. Kumar, M.; Khan, L.; Khan, M.Z.; Alhussan, A.A. AGF-HAM: Adaptive Gated Fusion Hierarchical Attention Model for Explainable Sentiment Analysis. *Mathematics* **2025**, *13*, 3892. [\[CrossRef\]](#)
6. Wu, Y.; Mi, Q.; Gao, T. A Comprehensive Review of Multimodal Emotion Recognition: Techniques, Challenges, and Future Directions. *Biomimetics* **2025**, *10*, 418. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Cambria, E.; Liu, Q.; Decherchi, S.; Xing, F.; Kwok, K. SenticNet 7: A Commonsense-based Neurosymbolic AI Framework for Explainable Sentiment Analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022*; European Language Resources Association: Paris, France, 2022; pp. 3829–3839.
8. Du, Z.; Wang, Y.; Yang, X.; Im, S.K.; Wang, Z. MISA-GMC: An Enhanced Multimodal Sentiment Analysis Framework with Gated Fusion and Momentum Contrastive Modality Relationship Modeling. *Mathematics* **2026**, *14*, 115. [\[CrossRef\]](#)
9. Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; González, F.A. Gated Multimodal Units for Information Fusion. In *Proceedings of the ICLR Workshop, Toulon, France, 24–26 April 2017*.
10. Arevalo, J.; Solorio, T.; Montes-y-Gomez, M.; González, F.A. Gated Multimodal Networks. *Neural Comput. Appl.* **2020**, *32*, 10209–10228. [\[CrossRef\]](#)
11. Tsai, Y.H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.P.; Salakhutdinov, R. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*; Association for Computational Linguistics: Kerrville, TX, USA, 2019.
12. Moorthy, S.; Moon, Y.K. Hybrid Multi-Attention Network for Audio–Visual Emotion Recognition Through Multimodal Feature Fusion. *Mathematics* **2025**, *13*, 1100. [\[CrossRef\]](#)
13. Li, F.; Zhang, D. Transformer-Driven Affective State Recognition from Wearable Physiological Data in Everyday Contexts. *Sensors* **2025**, *25*, 761. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017*.
15. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2020**, arXiv:1910.01108v4.
16. Moreno Novoa, M.; Puertas, E.; Martinez-Santos, J.C. UTBNLP at SemEval-2025 Task 11: Predicting Emotion Intensity with BERT and VAD-Informed Attention. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), Vienna, Austria, 31 July–1 August 2025*; Association for Computational Linguistics: Kerrville, TX, USA, 2025; pp. 1217–1222.
17. Poria, S.; Cambria, E.; Bajpai, R.; Hussain, A. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Kerrville, TX, USA, 2017.
18. Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; Cambria, E. DialogueRNN: An attentive RNN for emotion detection in conversations. *AAAI Conf. Artif. Intell.* **2019**, *33*, 6818–6825. [\[CrossRef\]](#)
19. Sarkar, S. Sentiment Analysis for Mental Health. Kaggle Dataset. 2023. Available online: <https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health> (accessed on 29 April 2026).
20. Kallstenius, T.; Capusan, A.J.; Andersson, G.; Williamson, A. Comparing traditional natural language processing and large language models for mental health status classification: A multi-model evaluation. *Sci. Rep.* **2025**, *15*, 24102. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Demszky, D.; Movshovitz-Attias, D.; Ko, J.; Cowen, A.; Nemade, G.; Ravi, S. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020*; Association for Computational Linguistics: Kerrville, TX, USA, 2020; pp. 4040–4054. [\[CrossRef\]](#)
22. RobroKools. VAD-BERT. Hugging Face Model Card. BERT-Based Model Fine-Tuned to Predict Valence, Arousal, and Dominance from Text. 2022. Available online: <https://huggingface.co/RobroKools/vad-bert> (accessed on 7 July 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.